Techniques of Water-Resources Investigations
of the United States Geological Survey

Chapter A1

SOME STATISTICAL TOOLS IN HYDROLOGY

By H. C. Riggs

Book 4

HYDROLOGIC ANALYSIS AND INTERPRETATION

UNITED STATES DEPARTMENT OF THE INTERIOR

CECIL D. ANDRUS, Secretary

GEOLOGICAL SURVEY

V. E. McKelvey, Director

First printing 1968

Second printing 1969

Third printing 1977

UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON : 1968

For sale by the Branch of Distribution, U.S. Geological Survey,
1200 South Eads Street, Arlington, VA 22202

PREFACE

The series of manuals on techniques describes procedures for planning and executing specialized work in water-resources investigations. The material is grouped under major headings called books and further subdivided into sections and chapters; section A of Book 4 is on statistical analysis.

The unit of publication, the chapter, is limited to a narrow field of subject matter. This format permits flexibility in revision and publication as the need arises.

Provisional drafts of chapters are distributed to field offices of the U.S. Geological Survey for their use. These drafts are subject to revision because of experience in use or because of advancement in knowledge, techniques, or equipment. After the technique described in a chapter is sufficiently developed, the chapter is published and is sold by the U.S. Geological Survey, 1200 South Eads Street, Arlington, VA 22202 (authorized agent of Superintendent of Documents, Government Printing Office).

CONTENTS

FIGURES

TABLES

SOME STATISTICAL TOOLS IN HYDROLOGY

By H. C. Riggs

Abstract

This chapter of "Techniques of Water-Resources Investigations" provides background material needed for understanding the statistical procedures most useful in hydrology; it furnishes detailed procedures, with examples, of regression analyses; it describes analysis of variance and covariance and discusses the characteristics of hydrologic data.

Introduction

As hydrologic analyses become more sophisticated, the proper design and interpretation of these analyses require a greater knowledge of statistical methods. In fact, two long-used hydrologic tools, the flood-frequency curve and the duration curve, require an understanding of the theory of statistics for proper evaluation. The more elaborate statistical methods are mathematical, but graphical methods are extremely useful and adequately accurate for many purposes, if made with an understanding of the underlying assumptions and if properly interpreted.

Until recent years most statistics texts emphasized procedures applicable to normally distributed data because the assumption of normality is appropriate to many types of biological and agricultural data. But much of the data used in hydrology either is not normally distributed or does not have a probability distribution at all. Most hydrologists learned statistics from texts or courses directed toward analysis of normally distributed data. Consequently some early hydrologic analyses were either incorrectly done or incorrectly interpreted.

This chapter of "Techniques of Water-Resources Investigations" provides the background material needed for understanding the statistical procedures most useful in hydrology. Although it starts with the basic concept of a distribution, many elementary details are omitted. The reader is assumed to have some familiarity with statistical terminology, computation procedures, and elementary probability such as would be obtained from a classroom course in statistics for engineers or from the U.S. Geological Survey correspondence course "Elementary Statistics in Hydrology."

Although theory is emphasized in this chapter, the treatment is intuitive rather than rigorous. Many practical approaches to graphical regression are given, and the pitfalls associated with computation of least-squares lines are located. The chapter concludes with a discussion of the statistical characteristics of hydrologic data.

Distributions

The concept of a population of objects having a distribution of sizes (or of some other characteristic) is basic to the statistical method. It is not possible to collect enough data to define a frequency distribution exactly, but the existence of a particular one can be proven to the desired degree of confidence by repeating an experiment many times.

Kendall (1952, p. 23) reported the results of a dice-tossing experiment in which 12 dice were tossed simultaneously and the number of sixes was recorded for each toss. The dice were tossed 4,096 times with the results shown in table 1. Also shown are the relative frequency computed from the experimental results and the theoretical relative frequency computed from the binomial distribution. The close agreement between the theoretical and experimental frequencies indicates that the binomial distribution is applicable to this problem.

Table 1.—Results of dice-tossing experiment

[After Kendall (1952)]

No. of sixes	Frequency	Relative frequency	Theoretical relative frequency
0	447	0.109	0.112
1	1,145	.280	.269
2	1,181	.288	.296
3	796	.194	.197
4	380	.093	.089
5	115	.028	.029
6	24	.006	.007
7 and over	8	.002	.001
Total	4,096	1.00	1.00

The binomial distribution is a discrete distribution, that is, it can take values only at specific points along a scale. In the dice-tossing experiment, it is possible to obtain an integer number of sixes only; there is no such thing as 5.5 or 3.2 sixes.

More commonly, a variable may take any value along a scale. Such a variable and its distribution are known as continuous. A variable may be classified as continuous if it can take any value along a scale even though the limitations of measurement restrict the observations to discrete values. This condition exists with most natural phenomena.

To aid in understanding a distribution, consider 1,000 tree-ring indices ranging in size from 2 to 240. If these are grouped by six-unit increments of size, a histogram, or frequency distribution, is obtained (fig. 1). The irregularity of the profile of this distribution is due to the small (in a statistical sense) number of indices used in its preparation. The greater the number used, the smoother would be the profile of the frequency distribution. If the number of observations approaches infinity and the size increment approaches zero, the enveloping line of the frequency distribution will approach a smooth curve. Then if the ordinate values are divided by a number such that the area under the curve becomes one, the resulting curve is a probability density curve, or probability distribution, such as figure 2. The process just described requires the additional assumption that the variable can take any value within the range, that the variable is continuous, not discrete.

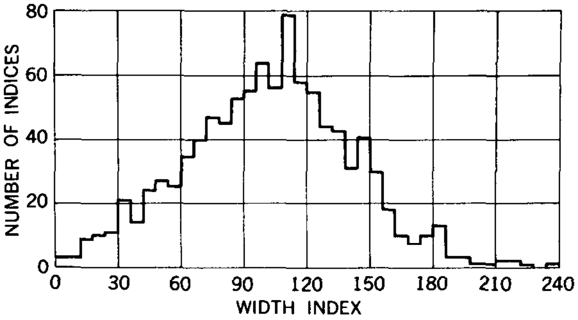

Figure 1.—Histogram, or frequency distribution, of 1,000 tree-ring indices.

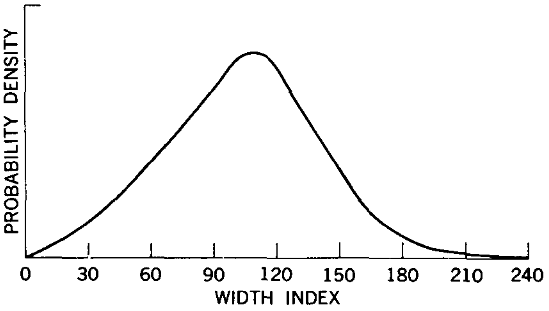

Figure 2.—Probability density curve of 1,000 tree-ring indices.

A theoretical probability distribution describes the relation between size (or some other other characteristic) and probability. For this relation to be valid, the individuals must occur randomly or be drawn randomly. The size of any individual drawn should not depend on the size of any one previously drawn. Probability, in the concept of frequency distributions, is defined as relative frequency. The distribution of the number of sixes obtained from repeated tosses of 12 dice may be illustrated by plotting the theoretical relative frequencies of table 1. The relative frequencies of each of the 6-unit increments in figure 1 could likewise be computed. In the first of these examples, a probability is associated with each possible outcome. In the second, a probability is associated with each increment of size; here the probability is of obtaining not a specific individual but any individual within the increment of size. This interpretation is required for continuous distributions because there is an infinite number of possible values and, thus, no probability of occurrence of a particular individual.

Referring again to figure 2, probability is related to the continuous distribution in the following way: The area under the curve represents the sum of all probabilities and therefore must equal one. Because every item was used in defining the distribution for which the total area is one, then the probability that any item will fall in the distribution is one and the probability that an item will fall in any segment of the distribution is the ratio of the area of that segment to the total area.

The distributions just described, both discrete and continuous, are called relative-frequency distributions, probability distributions, or just distributions. However, the probability interpretation is valid only if the data used are random. For example, the daily mean flow of a stream is closely related to the flows of previous days, so the distribution of daily means is not one to which the probability interpretation strictly applies. It is possible also to approximate a distribution which merely describes the sample. For instance, the distribution of grain sizes of a sample of a streambed is measured to characterize the material; there is no interest in the probability of obtaining a grain in a particular size range by additional sampling. Here the sample is not the individual grain but an aggregate of grains of various sizes.

Only a few standard theoretical distributions are widely used. Sampling theory and inference are based largely on the normal distribution with which the reader is assumed to be familiar. Other theoretical distributions will be introduced in this and succeeding chapters as appropriate.

Cumulative distributions

Suppose we know the probability density curve (probability distribution) for a variable and are interested in the probability of a random event being greater than some particular value E. This probability can be obtained by measuring or computing the proportion of the total area above the base value. For instance, the left curve of figure 3 shows an area under the curve to the right of E of 0.1, that is, $P=0.1$. Thus the probability of a random event exceeding E is 0.1.

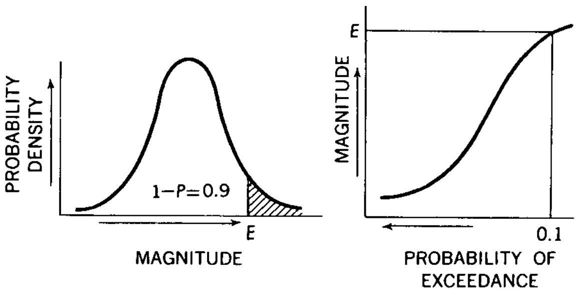

Figure 3.—Probability density curve (left) and its cumulative form (right).

Another form of the probability curve can be prepared by cumulating the probabilities from one end of the curve and plotting each of these cumulated probablities against the magnitude of its appropriate event. The cumulation is usually done mathematically. The result is the right-hand curve of figure 3. Cumulative distributions are commonly plotted to a probability scale such that the theoretical curve is a straight line. Such a scale can be devised for any two-parameter distribution. Normal probability plotting paper is widely known and used. Gumbel plotting paper is used in many hydrologic frequency analyses. (Although the Gumbel extreme-value distribution is a three-parameter distribution, one parameter, the skew, is constant for the form used and permits the construction of a scale which gives a straight-line plot). Both the normal and Gumbel probability plotting papers are available with either arithmetic or logarithmic ordinate scales. Thus plotting papers for four distributions—normal, log-normal, Gumbel, and log-Gumbel—are available.

When the probability density curve is cumulated from the right end, the probabilities of exceeding the various magnitudes are obtained. If cumulated from the left, probabilities of not exceeding those magnitudes are obtained. The appropriate cumulative curve, commonly called a frequency curve in hydrology, depends on the desired use.

The various theoretical cumulative distributions used in hydrology and methods of estimating their parameters from a sample of data are discussed in another chapter of this series.

Statistical inference

We have 54 years of record on the Rappahannock River of Virginia and might ask two questions about the mean flow. First, what is the mean flow for the period of record? This is a unique value which can easily be computed. The second question, what is the mean flow of the stream?, cannot be answered definitely. We can only assume that the mean of the 54-year sample is an estimate of the true (population) mean. In other words, we infer the population characteristics from those of a sample from that population.

Statistical inference is based on the theory of sampling. From a population of known characteristics many samples are drawn (either actually or conceptually), and the relation of the sample characteristics to the population characteristics is defined.

Sampling theory requires use of the concept of a probability distribution. Assume that the distribution of some random variable is normal with mean μ, and standard deviation, σ, as shown in figure 4. (The term "random," as used here, means that the probability of drawing any one item of the population is the same as for any other.)

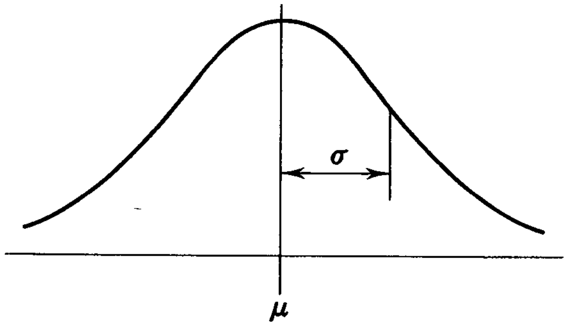

Figure 4.—Normal distribution.

Now suppose we take many samples of size N from this distribution, compute the mean of each of these samples, and compute the mean and variance of these sample means. The distribution of the means of samples of size N is superposed on the original distribution in figure 5. It can be shown that the distribution of the means is centered at μ and that the standard deviation of the distribution of means is $\sigma/\sqrt{N}$. Therefore, the mean of the means of samples of size N is an unbiased estimate of μ. Furthermore, the mean of one sample is an unbiased estimate of μ. Consequently, we infer that the sample mean, $\bar{X}$, is an estimate of the population mean. Obviously, if we used other samples we would obtain different estimates of the population mean.

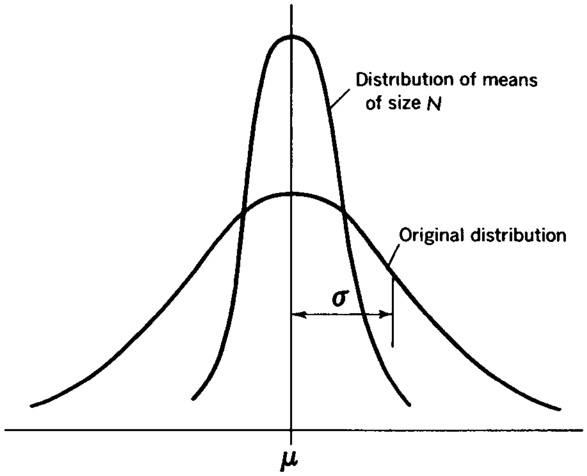

Figure 5.—Distribution of means of samples from a norma- distribution.

From a single sample, we can appraise the reliability of the estimate, $\bar{X}$, of the population mean. The distribution of means of values drawn from a normal distribution is normal. Consequently, two-thirds of the values should fall within one standard deviation ($\sigma/\sqrt{N}$) on each side of the mean. However, we do not know σ so we have to substitute S for it (where S is the standard deviation computed from the sample). The distribution of $\bar{X}$ having a standard deviation of $S/\sqrt{N}$ is known as the Student's t distribution, values of which are tabulated in statistics texts for various sizes of N.

Suppose now that we have K samples of size N and have defined K different sampling distributions of the mean of size N. For each sampling distribution, we can define a mean and a range of reliability, and we are intérested in whether such a range includes the true mean μ. Considering the range as a random interval, we may state that the probability (P) that the random interval includes μ is $1-e$, where e is the level of significance. Mathematically, for $e==0.32$,

$$P[(\bar{X}-\sigma/\sqrt{N})<\mu<(\bar{X}+\sigma/\sqrt{N})]=1-e=0.68.$$

The interval in brackets is called a confidence interval and the extremes are called confidence limits. Note that the above relation holds only if we use σ instead of S. If we use S, then $1-e$ is a function of the sample size, and the appropriate probability statement is

$$P[(\overline{X}-tS/\sqrt{N})<\mu<(\overline{X}+tS/\sqrt{N})]=1-e=0.68,$$

where Student's t is 1.09 for 10 degrees of freedom, for example. The width of the confidence interval increases as the level of significance decreases. For example, the 95-percent confidence limits ($e=0.05$) are

$$(\overline{X}-2.23S/\sqrt{N})<\mu<(\overline{X}+2.23S/\sqrt{N}),$$

where 2.23 is from the t table for 10 degrees of freedom.

The confidence interval described in the probability statement is a random interval, not a specific one. The probability statement (for $e=0.05$) means that 95 percent of a large number of intervals similarly obtained would include the true mean. This probability statement cannot be extended to one specific interval because a specific interval either contains the true mean or it does not and the probability is either one or zero. The true mean is not a variable, it is unique.

But we are interested in making a probability statement about one specific interval. We may say that the probability of our obtaining a random interval which includes the true mean is 0.95, or that we have 95-percent confidence that the interval obtained includes the true mean. Ordinarily in hydrologic reports it is only necessary to state the computed confidence interval and its level, not to interpret the meaning. See Mood (1950, p. 221–222) for a precise statement of the interpretation of a confidence interval.

Using the above theory, from a random sample, we can compute an estimate of the population mean and a measure of its reliability. This is an example of statistical inference.

Returning to the sampling theory, consider the distribution of variances of samples of size Λ from a normal distribution. This distribution is not centered around σ^2 but is to the left of it, as shown by the upper graph of figure 6. Therefore S^2 is known as a biased estimator of σ^2. It can be made unbiased by multiplying it by $N/(N-1)$ as shown in the lower sketch of figure 6. The standard deviation of the sampling distribution of $S^2(\Lambda/N-1)$ can also be computed.

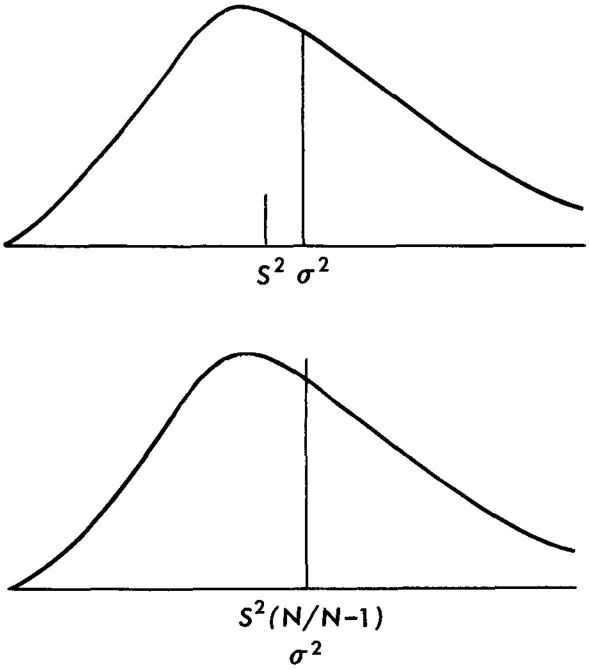

Figure 6.—Distribution of variances of samples.

A further use of inference is in testing hypotheses. One example will be given. Suppose we set up the null hypothesis, H_o, that the mean of a population is zero; that is, we hypothesize that there is no difference statistically between the mean and zero. This null hypothesis is written

$$H_o{:}\mu=0.$$

We draw a sample from this population and compute the statistics of the sampling distribution of the mean. We need some estimate of the hypothetical sampling distribution of means, and so we define it as normal with mean zero and standard deviation equal to $S/\sqrt{N}$ as computed from the sample (fig. 7).

Now if $\overline{X}$ (as computed from the sample) lies within one standard deviation of zero, we would conclude that there is no basis for doubting the hypothesis. If, on the other hand, $\overline{X}$ were two or three standard deviations away from zero, we would conclude that it is unlikely that the mean of the population is zero. For this

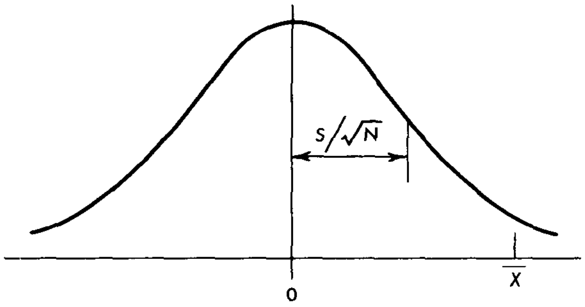

Figure 7.—Hypothetical sampling distribution of means.

latter condition the probability is small of obtaining an $\bar{X}$ of such size from a population having a mean of zero. Therefore, we would reject the hypothesis and would state that the result was significant at a certain probability level, meaning that the results obtained differ significantly from the hypothesis.

A common problem is the test of significance of a regression coefficient. The null hypothesis is again that the true value of the regression coefficient is zero, and the test may be made in the same way as before. However, the procedure commonly used is somewhat different. The confidence limits about the theoretical value are computed. If the regression coefficient is b, its standard error S_b, and its population value β, then the limits are found to be

$$b-tS_b<\beta<b+tS_b.$$

where t is the appropriate value for the chosen significance level and sample size. If the limits include zero, the hypothesis is accepted, that is, the regression coefficient is not significantly different from zero. If the limits are both on one side of zero, the hypothesis is rejected and the regression coefficient is considered significantly different from zero, that is, it is considered meaningful.

Many other tests of significance are available, but all parametric tests are based on the theory of sampling and follow the general procedure described above. A less powerful group of nonparametric tests may be used when the probability distribution of the statistic is not known. (See Siegel, 1956.)

Correlation and Regression

The distinctions between correlation and regression must be recognized in order to apply and interpret either of the methods. These distinctions are very marked although they may seem of little importance because of the similarity of the computation procedures. Dixon and Massey (1957, p. 189) made the following distinction between the two:

"A regression problem considers the frequency distribution of one variable when another is held fixed at each of several levels. A correlation problem considers the joint variation of two measurements, neither of which is restricted by the experiment."

Correlation is a process by which the degree of association between samples of two variables is defined. The correlation coefficient is a mathematical definition of that association. It is, of course, possible to compute a correlation coefficient from any two sets of data. The mathematical definition of association implies no cause-and-effect relation nor even that the relation between the two variables results from a common cause.

Correlation theory requires that the data be drawn randomly from a bivariate normal distribution. However, McDonald (1957) reported that experimental sampling studies show the nonnormality effects, usually regarded as disturbing by statisticians, to be of inconsequential magnitude geophysically. A further requirement of correlation is that both variables X and Y be without error due to measurement. Nothing can be measured without error, so the above requirement is one of degree. The question of the error allowable is subject to arbitrary decisions, particularly since the true error of the data is never known.

The end product of the process of correlation is the correlation coefficient; it is not an equation. The equations which describe Y as a function of X, and X as a function of Y, are regression equations, not correlation equations. Another way of stating the distinction between correlation and regression is that correlation measures the degree of association between two variables, whereas regression provides equations for estimating individual values of one variable from given values of the other.

Reliability of correlation results depends on the number of items used to compute the correlation coefficient and the magnitude of the computed correlation coefficient. Confidence limits are quite wide for samples of 30 items or less, unless the correlation coefficient is very large. For example, a chart shown by Bennett and Franklin (1954, p. 275) indicates that a correlation coefficient of +0.8 computed from a sample of 20 items would have a confidence belt extending from 0.6 to 0.9 for 95-percent probability. Because of this uncertainty, comparison of two correlation coefficients differing only by a few hundredths cannot be meaningfully interpreted. There also seems to be no justification for reporting correlation coefficients to more than two significant figures.

If the data can reasonably be assumed to be drawn from a normal bivariate distribution, then both correlation and regression analyses are appropriate. It is under this assumption that most of the examples in statistics texts are analyzed. However, regression is also appropriate under certain other conditions when correlation is not. The only assumptions required for regression are:

1. The deviations of the dependent variable about the regression line (for any fixed X) are normally distributed, and the same variance exists throughout the range of definition.
2. Values of the independent variable are known without error. The dependent variable is considered as an observation on a random variable, and the independent variable as some known constant associated with this random variable.
3. Observed values of the dependent variable are uncorrelated random events.
4. Each of the variables is homogeneous; that is, all individual values of a variable measure the same thing. Data are considered homogeneous if any subgroup to which certain of these data may be logically assigned has the same expected mean and variance as any other subgroup of the population. Neither variable need have a probability distribution in regression (but, of course, Y values corresponding to a fixed X are assumed to be normally distributed).

The end products of a regression analysis are two equations, $Y=f(X)$ and $X=f(Y)$ (usually only one is computed), because regression is directional. In contrast, correlation gives one index of the relation between variables.

The regression equation gives the average amount of change in the dependent variable corresponding to a unit change in the independent variable. Thus it gives more specific information than correlation. The regression coefficient can be tested to determine whether it is significantly different from zero, and this test is identical to the test of significance of the correlation coefficient (providing the data are drawn from a bivariate normal distribution).

The reliability of a regression is measured by the standard error, which is the standard deviation of the distribution (assumed normal) of residuals about the regression line (fig. 8 shows distribution of residuals). By definition, the standard error is the same throughout the range of X. This standard error was called the standard error of estimate by Ezekiel (1950, p. 131). It is also referred to as the standard error of regression and as the standard deviation from regression.

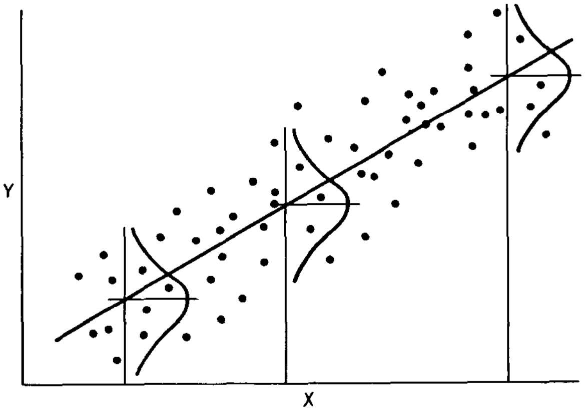

Figure 8.—Normal distribution of plotted points about the regression line.

The standard error of a prediction from regression is made up of three parts: the error of the mean, the error of the slope of the line, and the standard error of estimate. All three may be expressed in terms of the standard error of estimate so that the standard error of a prediction (S_p) is

$$S_p=S_e\sqrt{1+\frac{1}{n}+\frac{(X-\overline{X})^2}{\sum(X-\overline{X})^2}},$$

where S_e is standard error of estimate, n is number of items in the sample, and X is the independent variable. Thus the error of a prediction increases with distance from the mean (Snedecor, 1948, p. 120).

Most analyses require use of multiple correlation or regression. A multiple correlation is evaluated by partial correlation coefficients and by an index of total correlation. A partial correlation coefficient is an index of the degree of association between one independent variable and the dependent variable after the effects of the other independent variables have been removed.

In a multiple regression equation the regression coefficients are called partial regression coefficients. Each shows the effect on Y of a unit change in the particular independent variable, the effects of the other independent variables being held constant.

If the independent variables in a regression analysis are related to each other, the partial regression coefficients will be of a different magnitude from the simple regression coefficients. (The independent variables in a regression usually are related to each other as well as to the dependent variable.) See the section on "Application of the Regression Method" for elaboration on this subject (p .19).

The assumptions required for correlation are infrequently met in engineering problems and not generally met in hydrologic problems. Many of these problems to which the correlation method does not apply can be handled by the regression method because of the less restrictive assumptions. Thus the regression method may be used for such relations as that of concrete strength to time of setting, where neither value is randomly selected and neither variable has a probability distribution. Obviously the range of such a relation is limited to the range of the data selected.

Under the above conditions the correlation coefficient does not apply but, of course, can be computed from the relation

$$r=\sqrt{1-(S_e/S_y)^2},$$

where r=correlation coefficient, S_e=standard error of estimate, and S_y=standard deviation of the values of the dependent variable. From the above formula it can be seen that r depends on S_y, which depends on the range of data selected for problems such as the concrete strength relation to time of setting. Therefore, if the variables used in a regression are not randomly sampled, the computed value of r changes with the range of the arbitrarily selected sample and is therefore meaningless. Empirical verification of this statement is given by the data plotted in figure 9. (These data were selected to demonstrate this principle; the relation is not hydrologically significant.) Using all the points, the relation is computed to be

$$\log MAF=2.27+0.59 \log DA;$$

the standard error is 0.22 log unit and the computed correlation coefficient is 0.97. *MAF* is mean annual flood and *DA* is drainage area.

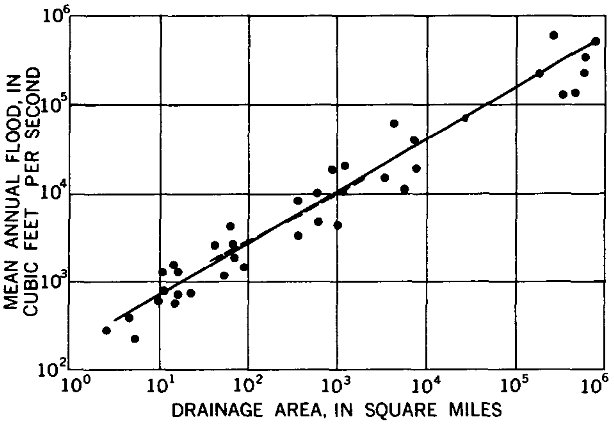

Figure 9.—Plot used in demonstrating the effect of sample range on computed correlation coefficient. Dashed line is the relation for 14 drainage areas ranging from 40 to 2,000 square miles.

If only the 14 points for drainage areas ranging from 40 to 2,000 square miles (fig. 9) are used, the relation is

$$\log MAF=2.31+0.57 \log DA.$$

This relation has a standard error of 0.23 log unit (almost the same as the previous standard error), but the computed correlation coefficient is 0.83, much lower than that obtained by using samples from a greater range. Obviously such variability in the correlation coefficient would render it unsuitable as a measure of the degree of relation for this type of application.

This manual emphasizes regression over correlation, not only because correlation is commonly inapplicable to particular hydrologic data but because regression provides quantitative answers to specific problems. In general, regression is preferred over correlation for hydrologic problems even when the data are suitable for a correlation analysis. Uses of regression analysis are:

1. To estimate individual values of the dependent variable corresponding to selected values of the independent variables.
2. To determine the amount of change in the dependent variable associated with a unit change in an independent variable.
3. To determine whether certain variables (which do not have probability distributions) are related to a dependent variable.
4. To improve estimates of the parameters defining the probability distribution of the dependent variable.

Correlation is most useful in theoretical studies and in time-series analysis.

Serial correlation

It has been pointed out that for a probability distribution to be valid the individuals must occur randomly or be drawn randomly. Hydrologic data such as daily stream discharges form a time series, that is, a sequence of values arranged in order of occurrence. The characteristics and analysis of hydrologic time series have been described by Dawdy and Matalas (1964). A common characteristic of a time series is the existence of a nonrandom element which produces a dependence between observations *k* units apart. This dependence is called serial correlation, and its degree is measured by the serial correlation coefficient.

First-order serial correlation is the dependence between observations adjacent in time; the *k*th order is the dependence between observations *k* units apart. A plot of the serial correlation coefficient against order is a correlogram (Dawdy and Matalas, 1964).

To determine the serial correlation, the time series is related to itself offset *k* units. For example, the time series in the first column below is related to itself shifted one observation to obtain the first-order serial correlation coefficient.

$$\begin{array}{ll} x_1 & — \\ x_2 & x_1 \\ x_3 & x_2 \\ . & . \\ . & . \\ . & . \\ . & . \\ x_n & x_n-1 \end{array}$$

Computational details are the same as for the relation between two variables and are given in the section on "Simple Linear Regression."

A test of significance of a first-order serial correlation coefficient is given by Dawdy and Matalas (1964).

Regression Methods

The previous section described regression in general terms and concluded with some uses of regression. This section describes the computation and interpretation of regression equations, both analytical and graphical, and some characteristics of the regression method.

Regression models

We begin a regression problem with a dependent variable which we want to predict from one or more independent variables. The independent variables are values or characteristics which seem to be physically related to the dependent variable. Next we need a model which describes the way in which the independent variables are related to the dependent variable. The model should be in accord with known physical principles, but its exact form may be dictated by the data used.

Using a dependent variable, *Y*, and independent variables, *X* and *Z*, the equations and graphs of some more common regression models are shown in figure 10. Joint relations, those which include a variable which is the product of two other variables, have been discussed in detail by Ezekiel and Fox (1959). The product of two variables is called an interaction term. Combinations of the models shown in figure 10 can be used to describe more complicated relations, and the equations can be readily

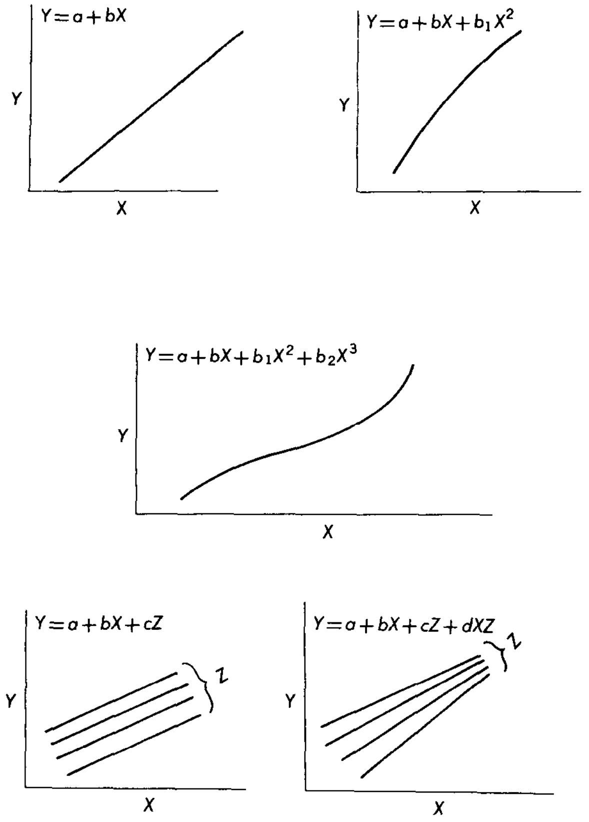

Figure 10.—Equations and graphs of some common regression models.

extended to include additional independent variables.

Having selected a suitable model, the coefficients in that model equation are computed from sample data by the method of least squares as described subsequently.

Note that although two of the graphs in figure 10 are curved, all of the model equations are in linear form. This linearity of the model equation is a requirement for direct least-squares solution. Linearity can sometimes be attained by transforming the variables.

Transformations

There are two principal reasons for transforming data before analysis: (1) to obtain a linear regression model, and (2) to achieve equal variance about the regression line throughout the range.

We have seen from figure 10 that certain two-variable regressions may be linearized without transforming the variables. The method is known as polynomial regression in which additional variables in successively higher powers of the independent variable are added to the model. But suppose we know or postulate that a relation should be of the form

$$Y=aX^b.$$

By taking logarithms of both sides of the equation the resulting linear equation is obtained:

$$\log Y=\log a+b \log X,$$

in which log a and b are constants which can be computed by a least-squares analysis using the variables log Y and log X. Likewise the relation

$$Y=ab^X$$

can be transformed to

$$\log Y= \log a+X \log b,$$

where log a and log b are the constants and log Y and X are the variables. Other transformations are sometimes used, but the logarithmic transformation is by far the most common.

The second reason for transforming data, and the more important one, is to achieve equal variance about the regression line. One of the assumptions basic to the regression method is that the distribution of errors about the regression line is normal and constant throughout the range (fig. 8). Again a log transformation is often used. For example, the graph on the left of figure 11 (from U.S. Geological Survey, 1949, p. 488) shows increasing scatter of points with increasing rainfall. But when the variables are plotted on the log chart (right graph, fig. 11), the scatter of points is almost uniform throughout the range. Thus, if it had been desired to carry the analysis beyond the graphical presentation, a transformation of the variables should have been made. (Ordinarily the variables would be reversed on the chart if a regression were to be made because runoff is the dependent variable.)

Other reasons for transforming data are to introduce additivity to the model and to achieve normality. The use of transformations is discussed by Acton (1959, p. 219–223).

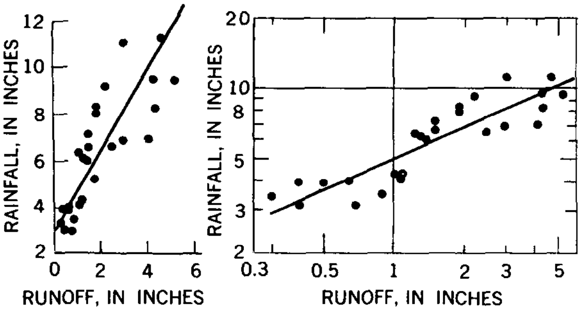

Figure 11.—Data from U.S. Geological Survey (1949, p. 488) plotted on natural and log scales showing the achievement of equal variance about the regression line by use of the log transformation.

Only the log transformation has been used in the above examples because it is by far the most common and useful. Other transformations such as the square root may be appropriate for certain data.

Table 2.—Data and computations for example of two-variable regression

Year	Runoff[1] (Y)	Precipitation[2] (X)	XY	X^2	Y^2
1928	125	110			
1929	67	73			
1930	68	74			
1931	71	91			
1932	118	108			
1933	144	130			
1934	169	152			
1935	138	134			
1936	102	98			
1937	91	90			
1938	125	119			
1939	87	77			
1940	84	100			
1941	58	84			
1942	79	85			
1943	124	115			
1944	62	70			
1945	87	91			
Σ	1,799	1,801	192,042	189,291	197,373
Mean	99.94	100.06			

[1] Annual runoff in percent of mean (Bumping River near Nile, Wash.).
[2] Annual rainfall in percent of mean (at Bumping Lake, Wash.).

Simple linear regression

Computation of a regression equation using the model $Y=a+bX$ is demonstrated using the data given in table 2. That table also shows computations of means, cross products, and squares. The individual cross products and squares need not be recorded; the sum of cross products, or squares, can be cumulated on a desk calculator. Such calculations are ordinarily checked by repeating the operation. The coefficients a and b in the regression equation, and the standard error of estimate are computed as shown below.

$$b=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sum X^2-\frac{(\sum X)^2}{N}}=\frac{\sum XY-N\overline{X}\overline{Y}}{\sum X^2-N\overline{X}^2},$$

$$b=\frac{192{,}042-\frac{(1{,}801)(1{,}799)}{18}}{189{,}291-\frac{(1{,}801)^2}{18}}=1.325.$$

Regression coefficient

$$a=\overline{Y}-b\overline{X}=99.94-(1.325)(100.06)=-32.6.$$

Intercept

Then

$$Y=a+bX=-32.6+1.32X,$$

or

$$Y=\overline{Y}+b(X-\overline{X})=99.94+(1.325)(X-100.06),$$

$$Y=-32.6+1.32X.$$

Equation of least-squares line

$$S_x^2=\frac{\sum X^2-\frac{(\sum X)^2}{N}}{N-1}=\frac{189{,}291-\frac{(1{,}801)^2}{18}}{17}$$

$$=534.76.\quad \text{Variance of } X$$

$$S_y^2=\frac{\sum Y^2-\frac{(\sum Y)^2}{N}}{N-1}=\frac{197{,}373-\frac{(1{,}799)^2}{18}}{17}$$

$$=1{,}033.71.\quad \text{Variance of } Y$$

$$S_{y\cdot x}^2=\frac{N-1}{N-2}\,[S_y^2-b^2S_x^2]=\frac{17}{16}[1{,}033.71$$

$$-(1.325)^2(534.76)]=100.8.$$

$$S_{y\cdot x}=10.0.\qquad \text{Standard error of estimate of } Y$$

$$r=\frac{bS_x}{S_y}=(1.325)\left(\frac{23.13}{32.15}\right)=0.95.$$

Correlation coefficient

The regression coefficient can be tested for significance as follows (Bennett and Franklin, 1954, p. 228):

$$S_b^2=\frac{S_{y\ x}^2}{\sum(x^2)}=\frac{100.8}{189{,}291-(1{,}801)^2/18}=0.011$$

Testing the hypothesis that $\beta=0$,

$$t_{n-2}=\frac{b-\beta}{S_b}=\frac{1.325-0}{0.105}=12.6$$

From a table of t, $t_{16,\ 0.01}=2.92$; therefore b is significantly different from zero. The 99-percent confidence limits for β are

$$1.325-2.92(0.105)<\beta<1.325+2.92(0.105)$$

or

$$1.02<\beta<1.63$$

The locus of the regression equation and the data used are shown in figure 12.

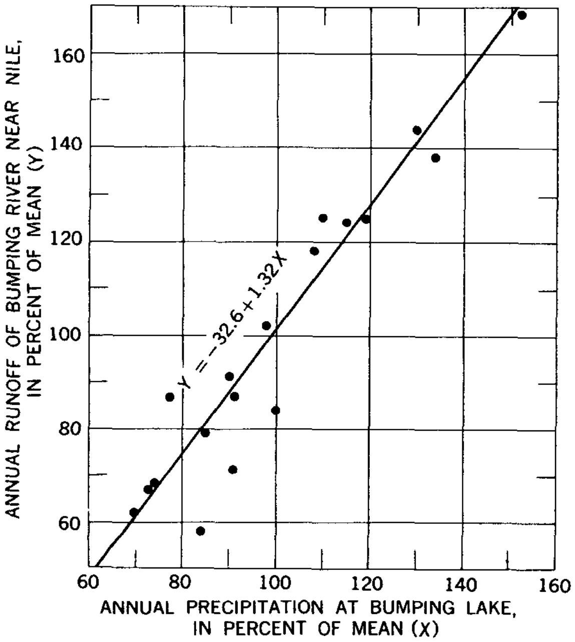

Figure 12.—Plot of data from table 2 showing computed regression line.

Another example showing the detailed computation of a regression equation is given by Ezekiel and Fox (1959, p. 57–63).

Multiple linear regression

The regression constants in a multiple linear regression model are computed from normal equations. For two independent variables the normal equations are

$$\sum(x_2^2)b_2+\sum(x_2x_3)b_3=\sum(x_1x_2),$$
$$\sum(x_2x_3)b_2+\sum(x_3^2)b_3=\sum(x_1x_3),$$

and

$$a=\overline{X}_1-b_2\overline{X}_2-b_3\overline{X}_3,$$

where the symbol $\overline{X}_i$ represents the mean of the ith variable, X_i represents a particular value of the ith variable, and x_i represents $(X_i-\overline{X}_i)$, the deviation from the mean of that variable. It is simpler to compute the squares and cross products of the variables in terms of X and then convert the results in terms of x than to begin with deviations from the mean. The conversion equations are

$$\sum(x_1x_2)=\sum(X_1X_2)-N\overline{X}_1\overline{X}_2,$$
$$\sum(x_2^2)=\sum(X_2^2)-N\overline{X}_2^2,$$
$$\sum(x_1x_3)=\sum(X_1X_3)-N\overline{X}_1\overline{X}_3,$$
$$\sum(x_2x_3)=\sum(X_2X_3)-N\overline{X}_2\overline{X}_3,$$

and

$$\sum(x_3^2)=\sum X_3^2-N\overline{X}_3^2,$$

where the last term in each equation is called the correction item and N is the number of items in the sample. In this notation X_1 is the dependent variable.

For three independent variables the normal equations are

$$\sum(x_2^2)b_2+\sum(x_2x_3)b_3+\sum(x_2x_4)b_4=\sum(x_1x_2),$$
$$\sum(x_2x_3)b_2+\sum(x_3^2)b_3+\sum(x_3x_4)b_4=\sum(x_1x_3),$$
$$\sum(x_2x_4)b_2+\sum(x_3x_4)b_3+\sum(x_4^2)b_4=\sum(x_1x_4),$$

and

$$a=\overline{X}_1-b_2\overline{X}_2-b_3\overline{X}_3-b_4\overline{X}_4,$$

where the symbols are the same as before.

The method of computation is best described by use of an example. The model, the data, and the preliminary computations are shown in table 3. Note that the logs of the original values are the variables being related. The need for this transformation was indicated by a preliminary graphical analysis. Only the cumulative sums of cross products and squares are taken from the calculating machine and recorded in table 3; individual values are not needed. Calculations are carried to five figures behind the decimal point when the variables are logarithms because the converted sums may be small relative to the numbers being subtracted from each other. The correction items shown in table 3 are obtained from the last term in the appropriate conversion equation. For X_2X_3, the appropriate equation is

$$\sum(x_2x_3)=\sum(X_2X_3)-N\overline{X}_2\overline{X}_3,$$

and the correction item of table 3 is $N\overline{X}_2\overline{X}_3$.

Subtracting the correction item from $\Sigma(X_2X_3)$ gives $\Sigma(x_2x_3)$, which is the corrected sum in table 3. This and the other corrected sums are

Table 3.—Multiple regression example: Tennessee low-flow characteristics

[Model is Log $Q_{20}=\log a + b_2 \log Q_2 + b_3 \log A + b_4 \log S$]

Station	Discharge at 2-year recurrence interval (cfs)	Drainage area (sq mi)	Slope of base-flow recession curve (percent)	Discharge at 20-year recurrence interval (cfs)	Log Q_2	Log A	Log S	Log Q_{20}										
	Q_2	A	S	Q_{20}	X_2	X_3	X_4	X_1	X_2X_3	X_2X_4	X_1X_2	X_3^2	X_3X_4	X_1X_3	X_4^2	X_1X_4	X_1^2	X_2^2
Hatchie-Stanton	330	1940	83	240	2.51851	3.28780	1.91908	2.38021										
Hatchie-Bolivar	184	1430	76	86	2.26482	3.15534	1.88081	1.93450										
Wolf-Rossville	150	503	89	100	2.17609	2.70157	1.94939	2.00000										
S. F. Forked Deer-Jackson	96	574	88	64	1.98227	2.75891	1.94448	1.80618										
S. F. Forked Deer-Chestnut Bluff	153	1100	86	107	2.18469	3.04139	1.93450	2.02938										
M. F. Forked Deer-Alamo	93	410	93	65	1.96848	2.61278	1.96848	1.81291										
Obion-Obion	315	1880	86	210	2.49831	3.27416	1.93450	2.32222										
Rutherford Fork Obion-Bradford	21.5	203	86	13	1.33244	2.30750	1.93450	1.11394										
S. F. Obion-Greenfield	95	431	94	66	1.97772	2.63448	1.97313	1.81954										
N. F. Obion-Union City	99	490	90	82	1.99564	2.69020	1.95424	1.91381										
Barren-Trousdale	45	132	87	31	1.65321	2.12057	1.93952	1.49136										
Buffalo-Flatwoods	138	447	81	85	2.13988	2.65031	1.90849	1.92942										
Buffalo-Lobelville	223	707	80	140	2.34830	2.84942	1.90309	2.14613										
Big Sandy-Bruceton	48	205	89	25.5	1.68124	2.31175	1.94939	1.40654										
Calfkiller-Sparta	25	178	84	13.5	1.39794	2.25042	1.92428	1.13033										
Clear Fork-Robbins	4.5	278	25	.4	.65321	2.44404	1.39794	−.39794										
Clinch-Tazewell	212	1474	75	108	2.32634	3.16850	1.87506	2.03342										
Collins-McMinnville	83	624	70	46	1.91908	2.79518	1.84510	1.66276										
Cypress-Florence	61	209	90	47	1.78533	2.32015	1.95424	1.67210										
Doe-Elizabethon	57	137	78	28	1.75587	2.13672	1.89209	1.44716										
Duck-Manchester	16.5	107	75	9	1.21748	2.02938	1.87506	.95424										
Elk-Estill Springs	55	282	87	36	1.74036	2.45025	1.93952	1.55630										
Falling Water-Cookeville	4.1	73.3	80	1.9	.61278	1.86510	1.90309	.27875										
Flint-Chase, Ala	82	342	85	59	1.91381	2.53403	1.92942	1.77085										
French Broad-Newport	850	1858	78	370	2.92942	3.26905	1.89209	2.56820										
L. Pigeon-Sevierville	68	353	72	26	1.83251	2.54777	1.85733	1.41497										
Nolichucky-Embreeville	355	805	60	140	2.55023	2.90580	1.77815	2.14613										
Piney-Vernon	62	193	85	38	1.79239	2.28556	1.92942	1.57978										
Richland-Pulaski	23.7	366	67	10	1.37475	2.56348	1.82607	1.00000										
Powell-Arthur	109	685	76	64	2.03743	2.83569	1.88081	1.80618										
Red-Adams	61	678	85	33	1.78533	2.83123	1.92942	1.51851										
Roaring-Hilham	6	70.8	80	2.3	.77815	1.85003	1.90309	.36173										
Shoal-Iron City	100	348	68	52	2.00000	2.54158	1.83251	1.71600										
Sequatchie-Whitwell	48	384	74	24	1.68124	2.58433	1.86923	1.38021										
S. Chickamauga-Chickamauga	105	428	78	73	2.02119	2.63144	1.89209	1.86332										
Sewee-Decatur	19	117	79	10	1.27875	2.06819	1.89763	1.00000										
Stones-Smyrna	15	552	59	4	1.17609	2.74194	1.77085	.60206										
Toccoa-Dial, Ga	150	177	85	110	2.17609	2.24797	1.92942	2.04139										
Turtletown-Turtletown	18	26.9	88	10	1.25527	1.42975	1.94448	1.00000										
Tellico-Tellico Plains	41	118	74	22	1.61278	2.07188	1.86923	1.34242										
Sums					72.32542	101.79564	75.63122	61.55501	190.44213	137.37714	123.04731	265.96512	192.40368	163.22578	143.33712	117.54964	109.03684	140.97663
Means					1.80814	2.54489	1.89078	1.53888										
Correction items									184.06080	136.75160	111.30040	259.05880	192.47320	156.65120	143.00200	116.38720	94.72600	130.77480
Corrected sums									6.38133	.62554	11.74691	6.90632	−.06952	6.57458	.33512	1.16244	14.31084	10.21083

N = 40.

then substituted in the normal equations. The computation of regression coefficients is shown below with the explanation following.

$N=40$

Normal equations (see Ezekiel, 1950, p. 198):

$$
\begin{array}{rll}
 & \text{I} & \sum(x_2^2)b_2+\sum(x_2x_3)b_3+\sum(x_2x_4)b_4=\sum(x_1x_2) \\
 & \text{II} & \sum(x_2x_3)b_2+\sum(x_3^2)b_3+\sum(x_3x_4)b_4=\sum(x_1x_3) \\
 & \text{III} & \sum(x_2x_4)b_2+\sum(x_3x_4)b_3+\sum(x_4^2)b_4=\sum(x_1x_4) \\
 & \text{I} & 10.20183b_2+6.38133b_3+0.62554b_4=11.74691 \\
 & \text{I}' & -b_2-0.625508b_3-0.061316b_4=-1.15145 \\
 & \text{II} & 6.38133b_2+6.90632b_3-0.06952b_4=6.57458 \\
(-0.625508) & \text{I} & -6.38133b_2-3.99157b_3-0.39128b_4=-7.34779 \\
 & \sum_2 & 2.91475b_3-0.46080b_4=-0.77321 \\
 & \text{II}' & -b_3+0.158092b_4=0.26527 \\
 & \text{III} & 0.62554b_2-0.06952b_3+0.33512b_4=1.16244 \\
(-0.061316) & \text{I} & -0.62554b_2-0.39128b_3-0.03836b_4=-0.72027 \\
(0.158092) & \sum_2 & 0.46080b_3-0.07285b_4=-0.12224 \\
 & \sum_3 & 0.22391b_4=0.31993 \\
 & & b_4=1.42883 \\
 & \text{II}' & -b_3+(0.158092)(1.42883)=0.26527 \\
 & & b_3=-0.03938 \\
 & \text{I}' & -b_2-(0.625508)(-0.03938)-(0.061316)(1.42883)=-1.15145 \\
 & & -b_2+0.02463-0.08761=-1.15145 \\
 & & b_2=1.08847 \\
 & \text{III} & (0.62454)(1.08847)-(0.06952)(-0.03938)+(0.33512)(1.42883)=1.16244 \\
 & & 1.16245\cong 1.16244 \quad \text{Check}
\end{array}
$$

The above computation utilizes the Doolittle method, a simplified method of solving simultaneous equations having a certain symmetry. The normal equations are on the first three lines. Next is the first normal equation with converted sums from table 3 substituted in it. Line 5 is obtained by dividing the equation next above by its coefficient of b_2 with the sign changed. Line 6 is the second normal equation, with converted sums from table 3 substituted in it. Line 7 is obtained by multiplying the equation of line 4 by the coefficient of b_3 in line 5. Line 8 is obtained by subtracting line 7 from line 6. Line 9 is line 8 divided by the coefficient of b_3 with the sign changed. Line 10 is the third normal equation. Line 11 is line 4 multiplied by the coefficient of b_4 on line 5 with the sign changed. Line 12 is line 8 multiplied by the coefficient of b_4 on line 9 with the sign changed. Line 13 is the sum of lines 10, 11, and 12. Lines 14–18 complete the computations of the regression coefficients. Lines 20 and 21 are used to check the results. Only the third normal equation provides a complete check.

The regression constant is obtained from

$$a=\bar{X}_1-b_2\bar{X}_2-b_3\bar{X}_3-b_4\bar{X}_4$$

$$a=1.53888-(1.08847)(1.80814)$$
$$-(-0.03938)(2.54489)-(1.42883)(1.89078)$$
$$a=-3.03061$$

Substituting the computed constants in the regression model gives

$$\log Q_{20}=-3.03+1.09\log Q_2$$
$$-0.04\log A+1.43\log S.$$

By taking antilogs this becomes

$$Q_{20}=0.00093Q_2^{1.09}A^{-0.04}S^{1.43}.$$

The standard error of estimate, S, is computed as follows

$$S^2=\frac{\sum(x_1^2)-b_2\sum(x_1x_2)-b_3\sum(x_1x_3)-b_4\sum(x_1x_4)}{N-M},$$

where N is the number of items in the sample and M is the number of lost degrees of freedom (one degree of freedom is lost for each constant in a regression equation). Substituting,

$$S^2=\frac{\begin{array}{l}14.31084-(1.08847)(11.74691)-\\(-0.03938)(6.57458)-(1.42883)(1.16244)\end{array}}{40-4},$$

$$\begin{aligned}S^2&=0.00341,\\S&=0.0584\\&=\text{standard error in log units.}\end{aligned}$$

The standard error of a regression having a logarithmic dependent variable is a constant percentage of the curve value throughout the range of Y rather than a constant magnitude in terms of the untransformed variable, as in the example of table 2.

To compute the standard error in percent look up the antilogs of $1+S$ and $1-S$. These antilogs are ratios to 10, from which the percentage deviation is obvious. Consider the standard error of 0.0584 log unit, computed above:

$$1+S=1.0584 \quad \text{Antilog } 11.4,$$

and

$$1-S=0.9416 \quad \text{Antilog } 8.75.$$

The percentage errors are

$$100(11.4-10)/10=+14 \text{ percent},$$

and

$$100(10-8.75)/10=-12.5 \text{ percent}.$$

The computation can be made very rapidly on a log-log slide rule.

A correlation coefficient is not computed for this problem because (1) the purpose of the problem is to get an estimating equation and (2) the data used cannot be considered as drawn from a multivariate normal distribution; therefore correlation is not appropriate and a computed correlation coefficient would have little meaning.

The standard error of estimate of this regression is a measure of its reliability and can be used to estimate the reliability of predictions made from the regression equation as described in the section on "Correlation and Regression." But the question may arise as to whether we might get as good a result using fewer variables, or whether each of the independent variables is related to the dependent variable. We could answer the first question by recomputing regression equations and standard errors using fewer variables, but to answer the second we need a test of significance of each regression coefficient. To make this significance test the regression needs to be computed somewhat differently, as described in the next section.

Regression computation using "c" multipliers

In this method the normal equations are expressed in terms of "c" multipliers rather than regression coefficients. The method affords two advantages, (1) significance tests of the regression coefficients are simply made, and (2) the regression equations for different dependent variables can be obtained from the same "c" multipliers. The method has been described by Ezekiel and Fox (1959, p. 499–503), Fisher (1950, p. 156–166), and Bennett and Franklin (1954, p. 248–255). The normal equations are

$$c_{22}\sum(x_2^2)+c_{23}\sum(x_2x_3)+c_{24}\sum(x_2x_4)=1,$$

$$c_{22}\sum(x_2x_3)+c_{23}\sum(x_3^2)+c_{24}\sum(x_3x_4)=0,$$

and

$$c_{22}\sum(x_2x_4)+c_{23}\sum(x_3x_4)+c_{24}\sum(x_4^2)=0.$$

From the above equations c_{22}, c_{23}, and c_{24} may be obtained. The additional elements needed, c_{32}, c_{33}, c_{34} and c_{42}, c_{43}, c_{44}, are obtained by solving similar equations with the right-hand sides replaced by 0, 1, 0 and 0, 0, 1, respectively.

The regression coefficients may then be evaluated by the equations

$$b_2=c_{22}\sum(x_1x_2)+c_{23}\sum(x_1x_3)+c_{24}\sum(x_1x_4),$$

$$b_3=c_{32}\sum(x_1x_2)+c_{33}\sum(x_1x_3)+c_{34}\sum(x_1x_4),$$

and

$$b_4=c_{42}\sum(x_1x_2)+c_{43}\sum(x_1x_3)+c_{44}\sum(x_1x_4),$$

where X_1 is the dependent variable.

To test the regression coefficients for significance, first compute the variance, S^2, of the observations X_1 about the regression surface. This variance is the square of the standard error of estimate and is obtained by the same formula used in the previous computation, that is,

$$S^2=\frac{\left[\begin{array}{l}\sum(x_1^2)-b_2\sum(x_1x_2)-b_3\sum(x_1x_3)\\ \qquad\qquad\qquad\qquad\qquad -b_4\sum(x_1x_4)\end{array}\right]}{N-M}$$

Then

$$\text{variance of } b_2=S^2(c_{22}),$$
$$\text{variance of } b_3=S^2(c_{33}),$$

and

$$\text{variance of } b_4=S^2(c_{44}),$$

and the standard errors are the square roots of the variances.

The confidence interval for β_2, a population regression coefficient, at a probability level α, may be expressed as

$$b_2-t_{N-4}S\sqrt{c_{22}}<\beta_2<b_2+t_{N-4}S\sqrt{c_{22}},$$

where t_{N-4} is from the t distribution with $N-4$ degrees of freedom at the selected α level.

The regression coefficient, b_2, is significantly different from zero if the confidence limits do not include zero.

The regression computation using "c" multipliers and the data of table 3 are given below. Solutions of the normal equations follow the same pattern as previously described. Solving for c_{22}, c_{23}, and c_{24}:

$$\begin{array}{rll}
 & \text{I} & \sum(x_2^2)c_{22}+\sum(x_2x_3)c_{23}+\sum(x_2x_4)c_{24}=1\\
 & \text{II} & \sum(x_2x_3)c_{22}+\sum(x_3^2)c_{23}+\sum(x_3x_4)c_{24}=0\\
 & \text{III} & \sum(x_2x_4)c_{22}+\sum(x_3x_4)c_{23}+\sum(x_4^2)c_{24}=0\\
 & \text{I} & 10.20183c_{22}+6.38133c_{23}+0.62554c_{24}=1\\
 & \text{I}' & -c_{22}-0.625508c_{23}-0.061316c_{24}=-0.0980216\\
 & \text{II} & 6.38133c_{22}+6.90632c_{23}-0.06952c_{24}=0\\
(-0.625508) & \text{I} & -6.38133c_{22}-3.99157c_{23}-0.39128c_{24}=-0.625508\\
 & \sum_2 & 2.91475c_{23}-0.46080c_{24}=-0.625508\\
 & \text{II}' & -c_{23}+0.158092c_{24}=0.214601\\
 & \text{III} & 0.62554c_{22}-0.06952c_{23}+0.33512c_{24}=0\\
(-0.061316) & \text{I} & -0.03836c_{24}=-0.061316\\
(0.158092) & \sum_2 & -0.07285c_{24}=-0.098888\\
 & \sum_3 & 0.22391c_{24}=-0.160204\\
 & & c_{24}=-0.71548\\
 & \text{II}' & -c_{23}+0.158092(-0.71548)=0.214601\\
 & & c_{23}=-0.32771\\
 & \text{I}' & -c_{22}-(0.625508)(-0.32771)-(0.061316)(-0.71548)=-0.0980216\\
 & & -c_{22}+0.204985+0.043870=-0.0980216\\
 & & c_{22}=0.34688\\
 & \text{III} & (0.62554)(0.34688)-(0.06952)(-0.32771)+(0.33512)(-0.71548)=0\\
 & & 0\cong 0 \quad \text{Check (to five places)}
\end{array}$$

Solving for c_{32}, c_{33}, and c_{34}:

$$\begin{array}{rll}
 & \text{I} & \sum(x_2^2)c_{32}+\sum(x_2x_3)c_{33}+\sum(x_2x_4)c_{34}=0\\
 & \text{II} & \sum(x_2x_3)c_{32}+\sum(x_3^2)c_{33}+\sum(x_3x_4)c_{34}=1\\
 & \text{III} & \sum(x_2x_4)c_{32}+\sum(x_3x_4)c_{33}+\sum(x_4^2)c_{34}=0\\
 & \text{I} & 10.20183c_{32}+6.38133c_{33}+0.62554c_{34}=0\\
 & \text{I}' & -c_{32}-0.625508c_{33}-0.061316c_{34}=0\\
 & \text{II} & 6.38133c_{32}+6.90632c_{33}-0.06952c_{34}=1\\
(-0.625508) & \text{I} & -6.38133c_{32}-3.99157c_{33}-0.39128c_{34}=0\\
 & \sum_2 & 2.91475c_{33}-0.46080c_{34}=1\\
 & \text{II}' & -c_{33}+0.158092c_{34}=-0.343082\\
 & \text{III} & .62554c_{32}-.06952c_{33}+0.33512c_{34}=0\\
(-0.061316) & \text{I} & -0.03836c_{34}=0\\
(0.158092) & \sum_2 & -0.07285c_{34}=0.158092\\
 & \sum_3 & 0.22391c_{34}=0.158092\\
 & & c_{34}=0.70605\\
 & \text{II}' & -c_{33}+0.158092(0.70605)=-0.343082\\
 & & c_{33}=0.45470
\end{array}$$

$$\text{I}'\quad -c_{32}-(0.625508)(0.45470)-(0.061316)(0.70605)=0$$

$$-c_{32}-0.28442-0.04329=0$$

$$c_{32}=-0.32771$$

$$\text{III}(0.62554)(-0.32771)-(0.06952)(0.45470)+(0.33512)(0.70605)=0$$

$$0=0 \qquad \text{Check (to five places)}$$

Solving for c_{42}, c_{43}, and c_{44}:

$$\begin{array}{rlrrr}
 & \text{I} & \sum(x_2^2)c_{42} & +\sum(x_2x_3)c_{43} & +\sum(x_2x_4)c_{44}=0 \\
 & \text{II} & \sum(x_2x_3)c_{42} & +\sum(x_3^2)c_{43} & +\sum(x_3x_4)c_{44}=0 \\
 & \text{III} & \sum(x_2x_4)c_{42} & +\sum(x_3x_4)c_{43} & +\sum(x_4^2)c_{44}=1 \\
 & \text{I} & 10.20183c_{42} & +\ 6.38133c_{43} & +\ 0.62554c_{44}=0 \\
 & \text{I}' & -c_{42} & -0.625508c_{43} & -0.061316c_{44}=0 \\
\hline
 & \text{II} & 6.38133c_{42} & +\ 6.90632c_{43} & -\ 0.06952c_{44}=0 \\
(-0.625508) & \text{I} & -6.38133c_{42} & -\ 3.99157c_{43} & -\ 0.39128c_{44}=0 \\
\hline
 & \sum_2 & & 2.91475c_{43} & -\ 0.46080c_{44}=0 \\
 & \text{II}' & & -c_{43} & +0.158092c_{44}=0 \\
 & \text{III} & 0.62554c_{42} & -\ 0.06952c_{43} & +\ 0.33512c_{44}=1 \\
(-0.061316) & \text{I} & & & -\ 0.03836c_{44}=0 \\
(0.158092) & \sum_2 & & & -\ 0.07285c_{44}=0 \\
\hline
 & \sum_3 & & & 0.22391c_{44}=1
\end{array}$$

$$c_{44}=4.46608$$

$$\text{II}'\qquad -c_{43}+0.158092(4.46608)=0$$

$$c_{43}=0.70605$$

$$\text{I}'\quad -c_{42}-(0.625508)(0.70605)-(0.061316)(4.46608)=0$$

$$-c_{42}-0.44164-0.27384=0$$

$$c_{42}=-0.71548$$

$$\text{III}\quad (0.62554)(-0.71548)-(0.06952)(0.70605)+(0.33512)(4.46608)=1$$

$$1.00003\cong 1 \qquad \text{Check}$$

Computing b coefficients and checking against those previously computed:

$b_2=c_{22}\sum(x_1x_2)+c_{23}\sum(x_1x_3)+c_{24}\sum(x_1x_4)$,

$=(0.34688)(11.74691)+(-0.32771)(6.57458)+(-0.71548)(1.16244)$,

$b_2=1.08851$ (1.08847 from previous computation; check).

$b_3=c_{32}\sum(x_1x_2)+c_{33}\sum(x_1x_3)+c_{34}\sum(x_1x_4)$,

$=(-0.32771)(11.74691)+(0.45470)(6.57458)+(0.70605)(1.16244)$,

$b_3=-0.03938$ (-0.03938 from previous computation; check).

$b_4=c_{42}(\sum(x_1x_2)+c_{43}\sum(x_1x_3)+c_{44}\sum(x_1x_4)$,

$=(-0.71548)(11.74691)+(0.70605)(6.57458)+(4.46608)(1.16244)$,

$b_4=1.42885$ (1.42883 from previous computation; check).

The coefficient a would be computed as described previously.

Computation of standard errors of b coefficients (Bennett and Franklin, 1954, p. 249):

$S_{1.234}=0.0584$ from previous computation.

$S_{b_2}=S_{1.231}\sqrt{c_{22}}=0.0584\sqrt{0.34688}=(0.0584)(0.589)$,

$=0.0344$ Standard error of b_2.

$S_{b_3}=S_{1.234}\sqrt{c_{33}}=0.0584\sqrt{0.4547}=(0.0584)(0.6743)$

$=0.0394$ Standard error of b_3.

$S_{b_4}=S_{1.234}\sqrt{c_{44}}=0.0584\sqrt{4.466}=(0.0584)(2.113)$,

$=0.1234$ Standard error of b_4.

Computation of confidence intervals of β coefficients (Bennett and Franklin, 1954, p. 250):

$t_{36,\ 0.95}=2.03$ (Dixon and Massey, 1957, table A–5, p. 384)

The table presented by Dixon and Massey (1957, table A–5, p. 384) gives values for one-half of the distribution. For 95-percent limits there would be 0.025 in each tail, and the value is taken in the column headed $t_{0.975}$. Notice that for infinite degrees of freedom the t and normal distributions are the same. In the normal distribution, 1.96σ on each side of the mean includes 95 percent of the items. In table A–5, 1.96 is listed under $t_{0.975}$.

The confidence limits are:

$$b_2-(t_{36,0.95})(S_{b2})<\beta_2<b_2+(t_{36,0.95})(S_{b2})$$
$$1.08851-(2.03)(0.0344)<\beta_2<1.08851+(2.03)(0.0344)$$
$$1.0187<\beta_2<1.1583$$
$$-0.03938-(2.03)(0.0394)<\beta_3<-0.03938+(2.03)(0.0394)$$
$$-0.1194<\beta_3<+0.0406$$
$$1.42885-(2.03)(0.1234)<\beta_4<1.42885+(2.03)(0.1234)$$
$$1.1783<\beta_4<1.6793$$

β is considered the true slope. Therefore the confidence limits give the range within which β lies with 95-percent probability. In this example the limits of β_3 include zero. This indicates that β_3 is not significantly different from zero at 95-percent level and that the parameter A should be eliminated from the regression.

Regressions having various numbers of independent variables

Examples have been given of computations for regressions having one and three independent variables, and the normal equations for a regression of two independent variables have also been given. The method of solution involving two variables is similar to that for three independent variables, but is much shorter.

Normal equations for regressions of four or more independent variables have been given by Ezekiel and Fox (1959, p. 181–183). Because computation of such regressions on a desk calculator is very time consuming, digital computers are being used.

Use of digital computers

Programs for regression computations are available for most computers, and regressions of more than two independent variables should ordinarily be made by digital computer rather than on a desk calculator. Simple regressions and regressions of two independent variables may be made quite rapidly on a desk calculator; use of a desk calculator for computations of these sizes may be advantageous.

Regression programs for digital computers vary but usually require listing of the data in floating decimal notation. These values are then punched on cards which are entered in the computer. Results are printed by the computer. A wide variety of options as to output is available. Detailed instructions for preparation of data and instructions to the computer should be obtained for the particular computer and program to be used.

Although no knowledge of regression analysis is necessary for preparing data for a computer program, some experience is needed to appraise the results. Opportunities for errors to be introduced into the process exist in the listing of the data and in its transferral to cards.

Questionable results may also be obtained if too few significant figures are carried through the computations. Only a person who has made regression computations the hard way can adequately judge whether the results of a regression analysis by digital computer (or any other method) are correct. The availability of digital computers has permitted ready computation of regressions using many variables, which has sometimes led to substitution of the computer for the analyst's brain. The problem should be solved by the analyst; the computer does the arithmetic.

Application of the regression method

An analytical problem to be solved by regression involves (1) selection of factors which are expected to influence the dependent variable, (2) describing these factors quantitatively, (3) selecting the regression model, (4) computing the regression equation, the standard error of estimate, and the significance of the regression coefficients, and (5) evaluating the results.

Selection of the appropriate factors should not be a statistical problem, but statistical concepts must enter into the process. If the analyst merely wants to know the relation of annual precipitation to annual runoff, he can proceed directly to selection of a model. But if his problem is to make the best possible estimate of runoff, he will include other factors, some of which may be related to each other as well as to runoff. The problem of determining if certain factors are related to the dependent variable requires careful selection of indices describing these factors quantitatively. These indices should accurately reflect the effects, and no two should describe the same thing. It is a characteristic of regression that if a factor is related to a dependent variable and this factor is entered in the regression model twice (as two different variables), the effect on the dependent variable will be divided equally between the two. Thus, if the total effect is small, the result of dividing it in two parts may be to produce nonsignificance in each of the parts. Likewise, several closely related variables may compute as nonsignificant, whereas one properly selected index would show a real effect. Thus, the independent variables should be selected with considerable care; the shotgun approach should not be used.

Another consideration in selection of variables is to avoid having a variable, or a part thereof, on both sides of the equation. Such a condition may be acceptable for certain problems, but the results must be evaluated carefully. A spurious relation may result, or the relation may be correct but its reliability difficult to assess. Benson (1965) described ways in which spurious relations may be built into a regression.

The user of the regression method should understand the effect of related independent variables on the computed regression coefficients. If the independent variables are entirely unrelated, the simple regression coefficients and the corresponding partial regression coefficients would be the same. However, such conditions rarely occur in nature. The multiple regression method provides a way of separating the total effect of the independent variables into the effect of each independent variable and an unexplained effect. Consider the simple regression

$$Y=a+b_1X_1\pm\text{error}, \qquad (1)$$

where Y also is affected by another variable, X_2, which is related to X_1. The regression using X_1 and X_2 will be

$$Y=a'+b_1'X_1+b_2X_2\pm\text{error}, \qquad (2)$$

where $b_1' \neq b_1$. If X_1 and X_2 are the only variables affecting Y (and the effects are linear), then equation 2 completely describes Y, and b_1' and b_2 are the true values of the regression coefficients (except for sampling errors). If X_1 and X_2 are positively correlated with each other and with Y, consider the effect on the magnitude of b_1. For each value of X_1 in equation 1, Y will appear to be more closely related than it actually is because X_2 increases with X_1 and its influence on Y is real though unmeasured. Therefore the regression coefficient b_1 is larger than its true value b_1.

Similar changes in b_1 and b_2 would occur if another factor, related to X_1 and X_2 and Y, were included in the regression. These changes in the magnitudes of the regression coefficients due to addition or deletion of a variable are characteristic of regression. They are sometimes interpreted as indicating that partial regression coefficients have no physical meaning. Such interpretations are not necessarily correct. If the variables used in the regression are selected on physical principles and the effects of each of the variables is appreciable, then the partial regression coefficients should be in accord with physical principles. In fact, it is good practice to compare the sign and the general magnitude of each partial regression coefficient with that expected. Benson (1962, p. 52–55) made a thorough comparison of this kind.

The regression coefficients of certain variables may change sign when another related variable is added to or deleted from the regression. This effect may result because (1) the variable is not a good index of the physical feature represented, (2) the effect of the variable is small relative to the sampling error, (3) the variable is so highly correlated with one or more other variables in the regression that the real effect is divided among them and no one variable shows a significant effect, and (4) the range of the variable sampled may be too small to define a significant effect.

A regression equation does not imply a cause-and-effect relation between the independent variables and the dependent variable. Both may be influenced by some other factor not readily measured. However, there should be some physical tie between the variables if the results can be considered meaningful.

Selection of a regression model usually begins with a graphical analysis. A model which plots as a straight line is commonly used unless there is strong evidence to the contrary.

If the sample data exist near an asymptote or near a maximum or minimum point on the curve, a simple model may be inadequate to describe the relation and a more sophisticated one may not be justified unless many data are available. An example showing the characteristics of three common models when applied to data defined near zero is given in figure 13. Physical considerations suggest that neither b nor Q_7 should be less than zero and that the line should be curved. The zero limitation can be obtained by using the variables log b and log Q_7 and thus making the curve asymptotic to zero on both axes. The addition of a term (log Q_7)^2 will provide the necessary curvature. The regression equation using these three variables is the top one on figure 13. It is not a good fit to the data.

Next, assume that it is not necessary that the curve be asymptotic to $Q_7=0$. Then a semilog model using the variables log b, Q_7, and Q_7^2 would be appropriate. But the equation based on this model reaches a maximum too soon and is a very poor fit. As a last resort assume a simple model with the variables b, Q_7, and Q_7^2. This equation is a good fit to the data, largely because of the locations of the data. An additional point of $b=0$ at $Q_7=10$ or more would have brought the curve below $b=0$ at $Q_7=7$. The curve shown on figure 13 reaches a minimum at $Q_7=7$ and increases beyond.

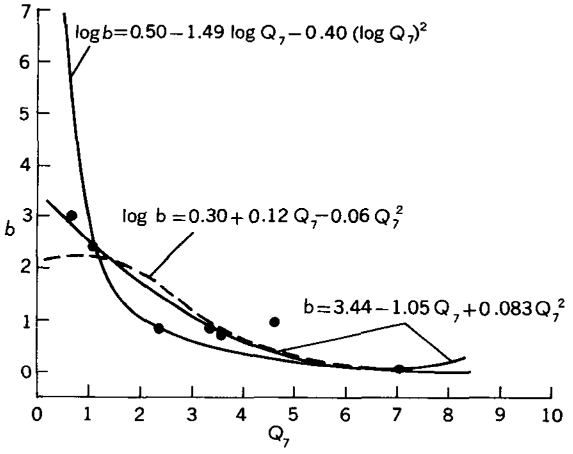

Figure 13.—Equations and graphs of three models based on the plotted data.

The mechanics of computing the regression equation, the standard error, and the tests of significance have been described. One important task remains, that of evaluating the results. First, the analyst should recognize that the regression equation developed, even though it is a good fit to the data, is not necessarily correct if extrapolated. For example, the curve corresponding to the bottom equation of figure 13 is a good fit to the seven points but increases directly with Q_7 for values of Q_7 greater than 7. On the other hand, the dashed curve of figure 13 fits the lower five points but becomes asymptotic to zero as Q_7 increases. Available information does not indicate which extrapolation is more nearly correct.

The signs of all significant regression coefficients should be in accord with physical principles. The regression is not necessarily incorrect if they are not; the nonconformity may be due to interrelations among the independent variables. Such a regression is useful for estimating values of the dependent variables from known values of the independent variables, and the reliability of the results, if within the defined range of the regression, can be computed.

The more difficult problem of determining whether a particular variable is related to the dependent variable may not have a definite answer. Even though a regression coefficient is statistically significant, there is a small probability that this result occurred by chance. Other samples could produce conflicting results. On the other hand, if many regressions produce nonsignificant coefficients of a particular variable, all coefficients having the same sign, then we would conclude that the effect of that variable was real but, of course, small.

A distinction should be made between statistical significance and practical significance. The regression coefficient of a variable may test highly significant, and yet the effect of that variable on the dependent variable may be negligible.

Uses and interpretations of regression analyses in hydrology have been discussed by Riggs (1960) and Amorocho and Hart (1964).

Graphical regression

The assumptions required of graphical regression are the same as those required for analytical regression. The results of a graphical regression can be expressed mathematically if no restrictions are added to the graphical analysis, and the standard error can be estimated.

Graphical regression is less restrictive than analytical regression in that the model need not be completely specified in advance. In fact, f an analytical model cannot be selected on a physical basis, it is conventional to prepare a preliminary graphical regression which will indicate an appropriate model. For example, consider the four data plots of figure 14. The first (upper left of fig. 14) indicates use of the model

$$Y=a+bX.$$

The second (upper right) requires

$$Y=a+bX+b_1X^2,$$

where the direction of curvature determines the sign of b_1. The third plot (lower left) indicates the need for a transformation unless the divergence can be explained by an additional variable. The fourth plot (lower right) shows no relation between Y and X, and, if only a two-variable relation is being considered, no further analysis would be made. A relation, however, between Y and X in the fourth plot may be obscured by the effect of another variable Z which has not been included. This aspect is discussed on page 23.

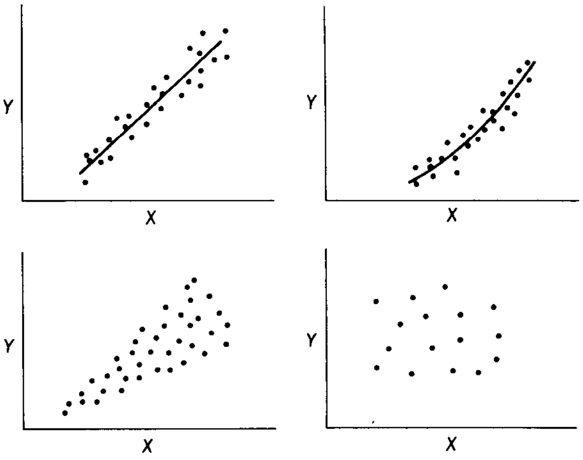

Figure 14.—Four possible outcomes of plotting Y against X.

The preparation of simple linear relations between two variables is well known. The regression line is not necessarily the same line as one would draw through the plotted points. There are two regression lines, one for $Y=f(X)$, and another for $X=f(Y)$ (fig. 15). The structural line, which balances the plotted points in both directions, has a slope approximately midway between the two regression lines. The differences in slope among the three lines depend on the degree of correlation of the variables. For perfect correlation all three lines have the same slope. Regardless of the correlation, both regression lines pass through the mean; the structural line may or may not pass through the mean.

To approximate the regression $Y=f(X)$, (1) group the points by small increments of X,

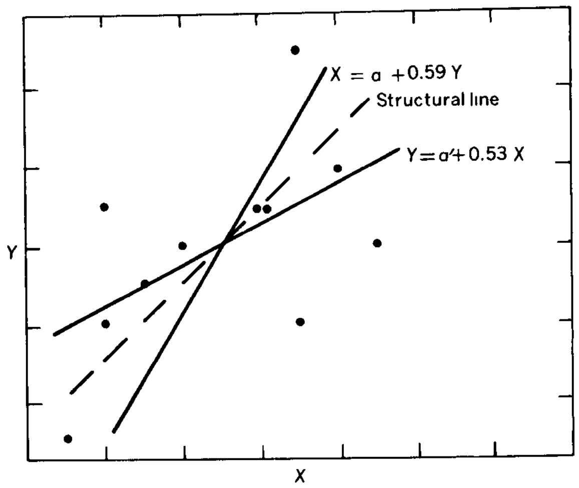

Figure 15.—Plot showing the two regression lines and the structural line.

(2) estimate the mean of each group in the Y direction, and (3) draw a line which averages these means. The procedure can be understood by referring to figure 15 and remembering that the distribution of points about the regression line in the Y direction is assumed to be the same throughout the range. Obviously that assumption cannot be true for a small number of points, but it is the condition which we try to approximate. The regression line of $Y=f(X)$ will have a flatter slope than that of a line drawn to balance the points in both Y and X directions.

The standard error of estimate of a graphical regression can be estimated readily. Remembering (1) that the standard error of estimate is the standard deviation of plotted points about the regression line, (2) that two-thirds of the points should be within one standard deviation on each side of the mean of a normal distribution, and (3) that a regression line theoretically passes through the mean value of Y corresponding to any value of X, then two lines, parallel to the regression line and one standard deviation above and below (in the Y direction), should encompass two-thirds of the plotted points. In practice it is simpler to draw the lines so as to exclude one-sixth of the points above and be ow, and then use the average of these two deviations from the mean as the estimated standard error of estimate. The procedure is illustrated in figure 16 for a log relation. The standard error can be described in log units but more common y is expressed in percent. This value is readily obtained by using div ders to lay off one standard error above and below a cycle separation if the relat on is plotted on log paper. The percentages are measured from one, as shown in figure 16.

For regressions on arithmetic plots, the standard error will be in the same units as Y and can be read from the plot.

The reliability of the graphically determ ned standard error is influenced by two factors having opposite effects. If the graphical-regression line has a steeper slope than the least-squares regression line, the graphical standard error will be la ger than the computed standard error. If we now assume that the graphical line of relation is the same as the least-squares line, the graphically computed standard error will underestimate the computed standard error when a few plotted points are far from the line but the majority are close. In any case the graphically determined standard error is only an approximation but is adequate for many problems.

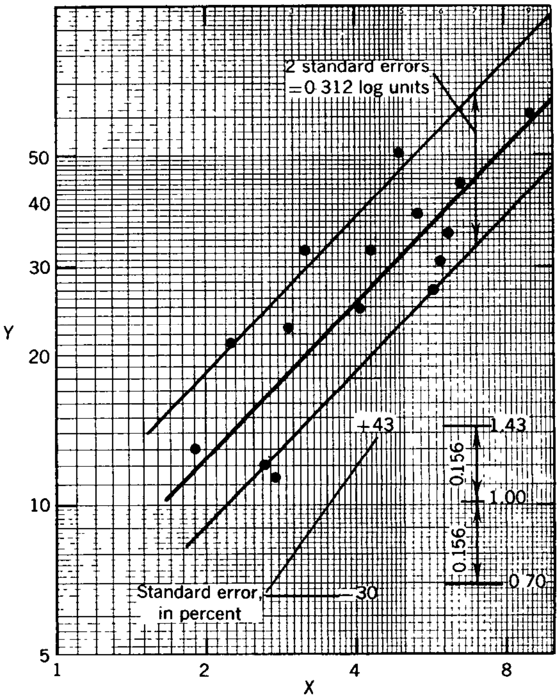

Figure 16.—Method of estimating the standard error of a graphical regression.

The correlation coefficient may also be estimated from a graphical regression by the relation

$$r=\sqrt{1-S_e^2/S_d^2},$$

where S_e is the graphically determined standard error and S_d is the standard deviation of the Y variables about their mean determined in the same manner as the standard error. Obviously, the correlation coefficient should be estimated only for relations between variables which can reasonably be assumed to be drawn from a bivariate normal distribution.

Graphical multiple regression

There are two general methods of graphical multiple regression. The method of deviations is based on the model

$$Y=a+b_1X_1+b_2X_2+\ .\ .\ .\ b_nX_n,$$

or a similar one allowing for curvilinearity. This method is probably the simplest and most useful one available.

The coaxial method of graphical multiple regression, used for runoff-precipitation relations, is a more flexible method than the method of deviations in that it allows both for interactions and curvilinearity. However, these advantages are obtained at the expense of much additional work and at the loss of a simple method of evaluating the reliability of the result. Linsley and others (1949, p. 650–655) described the procedure in detail. Unless stated otherwise, the descriptions of graphical multiple regression in this section refer to the method of deviations.

The purpose of multiple regression is to determine how a dependent variable changes with changes in two or more independent variables. This problem cannot be solved by considering one independent variable at a time because the independent variables are usually correlated to some extent with each other. This statement can be verified by analyzing the following synthetic data:

No.	*Y*	X_1	X_2
1	500	100	25
2	250	150	160
3	300	50	30
4	100	30	110
5	200	100	150
6	200	20	20
7	50	50	700

Assume that the logarithms of the variables are linearly related. This relation calls for plotting on log paper. First make a graphic comparison between *Y* and X_1 by plotting the appropriate data (see plot 1, fig. 17). (In statistical work the dependent variable is usually plotted on the ordinate scale.) Also plot *Y* against X_2 (plot 2, fig. 17). These plots indicate that *Y* cannot be estimated reliably from either parameter.

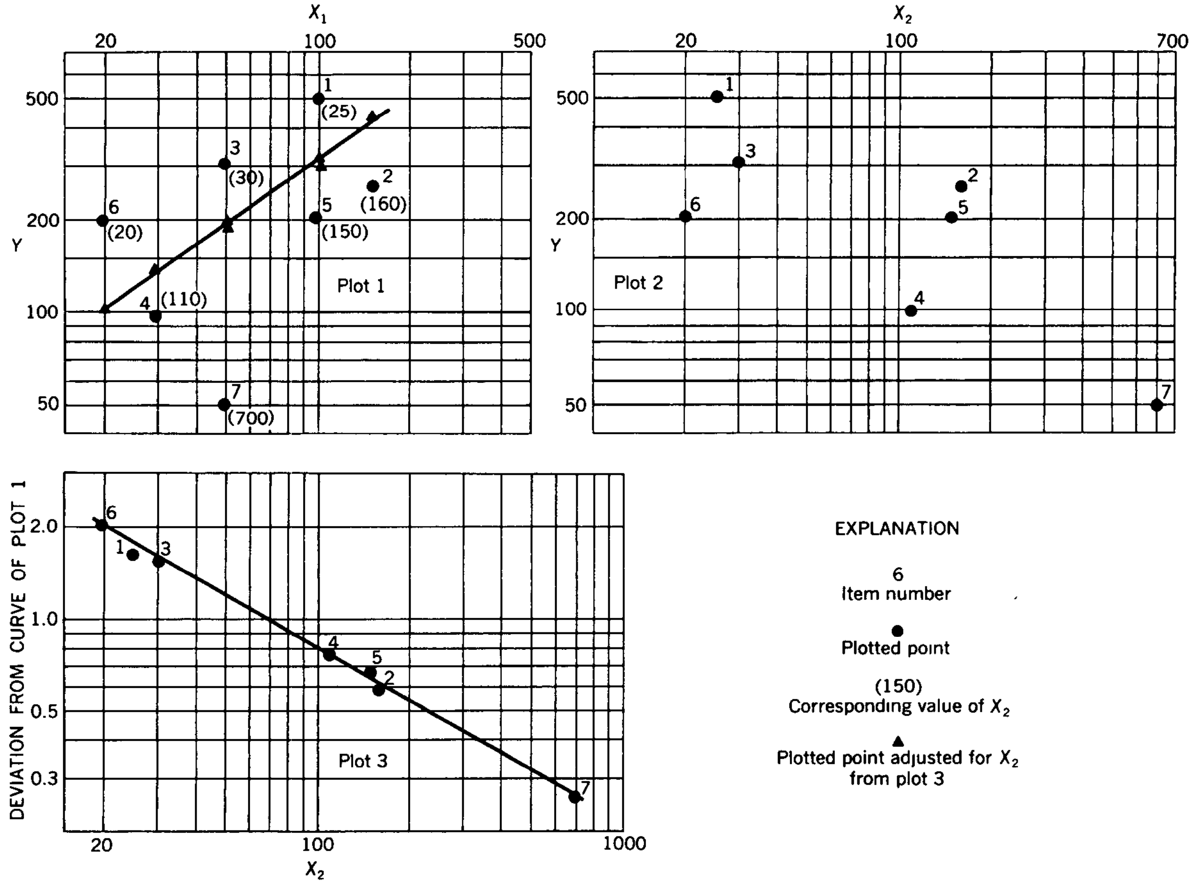

Figure 17.—Example of graphical multiple regression.

Now determine the relation between Y and both of the other variables. The procedure is as follows:

1. On figure 17, plot 1, write beside each point the corresponding value of X_2. It will be seen that the high values of X_2 tend to be on one side of the group and the low values on the other. This condition is an indication that X_2 values are related to Y. Draw a straight line through the points in such a way that it represents roughly some constant value of X_2. The line probably will not balance the plotted points.
2. Plot deviations (also called residuals) of Y from the straight line of plot 1 against X_2 as the abscissa on plot 3 (fig. 17). The deviations may be scaled from plot 1 or transferred by dividers. Because they are ratios, they should be measured above or below 1.00 on plot 3.
3. Draw a straight line averaging the points on plot 3.
4. Measure the deviations of the points from the curve of plot 3 and replot them on plot 1. These deviations are measured from the straight line in plot 1 and define the relation between Y and X_1 with the effect of X_2 removed. Sometimes these replotted points are not randomly distributed about the line, in which case the line should be redrawn and the whole process repeated. When a satisfactory balance is attained the regression is complete. The scatter of the adjusted points about the line of plot 1, is a measure of the error. The standard error of a graphical multiple regression may be approximated by using the adjusted points, as described in the section on "Graphical Regression." The line on plot 1 is the relation between Y and X_1 for the X_2 value at which the line of plot 3 crosses the 1.0 line (X_2=66). The relation of Y to X_1 for any other value of X_2 will be a line parallel to the line of plot 1, at a position defined by the curve of plot 3 for the desired value of X_2.

The example used gave much better results than ordinarily would be expected in hydrologic analyses. The data were manufactured (1) to illustrate the procedure and (2) to point out that a good relation may not be recognized if only two variables at a time are studied.

Graphical regressions involving more than two independent variables can be made. The residuals from each line are plotted against the next variable until all variables are used. Then the residuals from the last line are replotted from the first as described in step 4. In practical work it is usually difficult to define the effects of more than three independent variables, particularly when the influences of one or two of the variables are small.

Linear regression should be used whenever the plotted points do not definitely define a curve and when no physical reason is known for expecting the relation to be curved. If a curve or curves are indicated by both of the above criteria, then curves should be used. Complicated curves require four or more points for definition. They should be avoided when only a relatively few points are available to define the relation.

Graphical multiple regressions need not be made on logarithmic paper. Arithmetic plots can be handled as readily. Figure 18 relates summer runoff to spring water content of the snowpack and to summer precipitation. The graphical procedure is the same as in the first example, except that deviations are measured in the same units as the dependent variable and the deviation scale must have its center at zero with positive values above and negative values below. Obviously the mathematical model describing this relation would be different from one for a graphical relation developed on log paper.

The plotting paper selected for a particular problem should be that on which the distribution of the dependent variable for a fixed value of the independent variable is approximately the same for all values of the independent variable. This criterion is more important than that of attaining linearity.

Graphical multiple regression when the independent variables are highly correlated between themselves

Figure 19 demonstrates a technique that is sometimes useful in graphical regression. Data

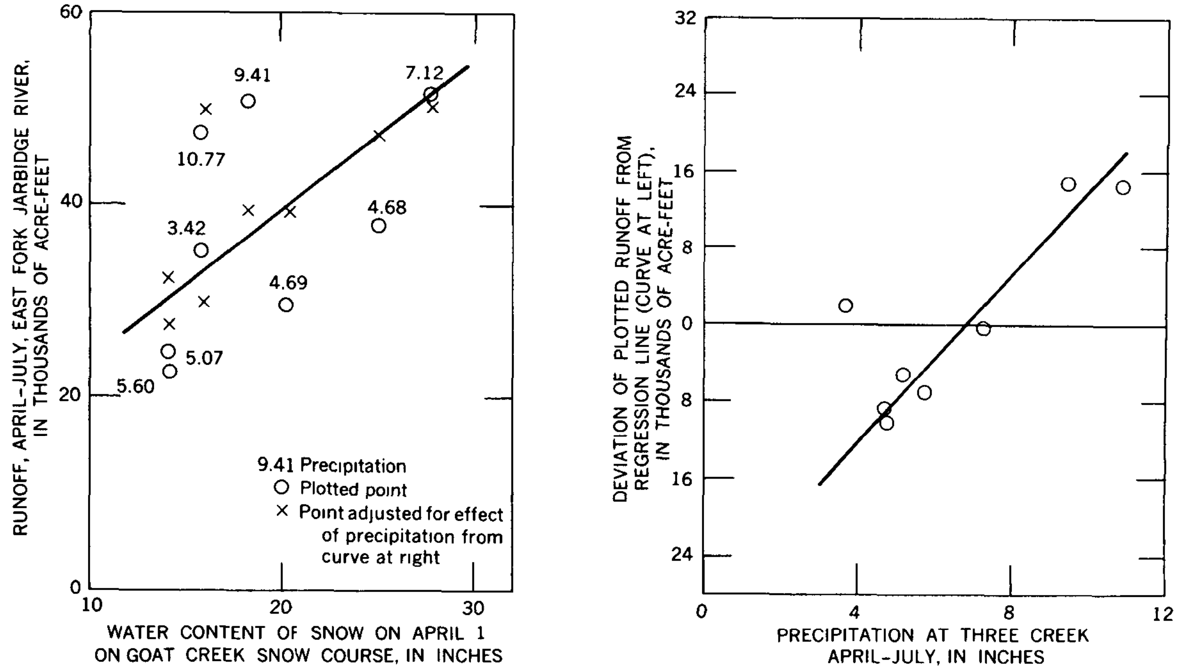

Figure 18.—Example of graphical multiple regression using arithmetic scales.

are given in table 4. Curve 1 (fig. 19) is the relation between 100-year flood (Q_{100}) and mean annual flood ($Q_{2.33}$) for 17 stations. The numerator of the fraction flagged to each plotted point is the mean annual discharge (Q_{av}) of the stream. Because this discharge

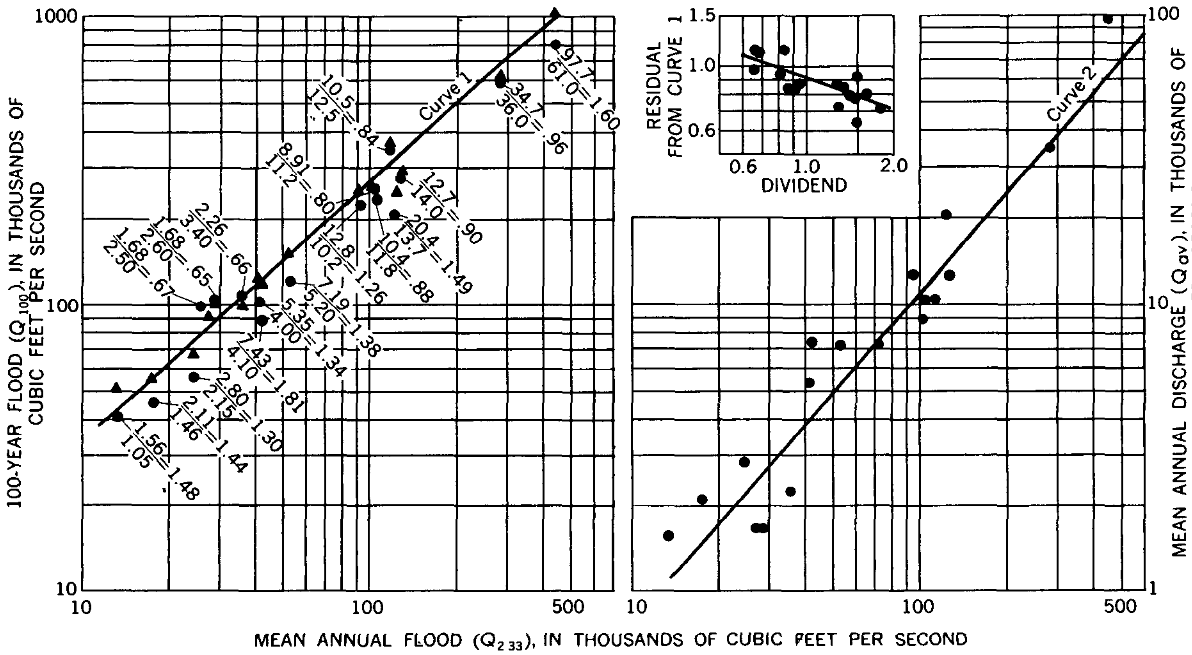

Figure 19.—Graphical regression using highly correlated independent variables. Based on data given in table 4.

Table 4.—Data for graphical regression using highly correlated independent variables
[From Riggs (1958)]

River and location	100-yr flood (cfs)	Mean annual flood (cfs)	Average discharge (cfs)
1. Neosho—Iola, Kans	105,000	28,300	1,680
2. Big Blue—Randolph, Kans	96,700	27,600	1,680
3. Miami—Dayton, Ohio	108,000	35,900	2,260
4. Savannah—Augusta, Ga	350,000	112,900	10,540
5. West Branch Susquehanna—Williamsport, Pa	260,000	104,100	8,910
6. Susquehanna—Towanda, Pa	236,000	107,200	10,370
7. Susquehanna—Harrisburg, Pa	594,000	282,600	34,700
8. Kanawha—Kanawha Falls, W. Va	276,000	125,600	12,670
9. Allegheny—Red House, N.Y	56,400	24,500	2,795
10. Iowa, Iowa City, Iowa	40,700	13,100	1,560
11. Tennessee—Knoxville, Tenn	228,000	94,700	12,820
12. French Broad—Asheville, N.C	45,500	17,500	2,112
13. Des Moines—Keosaugua, Iowa	103,000	41,100	5,351
14. Connecticut—White River Junction, Vt	122,000	53,100	7,190
15. Cumberland—Nashville, Tenn	208,000	122,200	20,400
16. Hudson—Mechanicville, N.Y	89,300	42,500	7,430
17. Ohio—Cincinnati, Ohio	800,000	443,700	97,700

increases with $Q_{2.33}$ it is impossible to tell by inspection whether use of Q_{av} will improve the relation.

The following procedure may be used to define the effect, if any, of Q_{av} on the scatter of points about curve 1:

1. Plot Q_{av} against $Q_{2.33}$ (as abscissa) and draw the mean line (curve 2).
2. Divide each Q_{av} by its value from curve 2 at the same value of $Q_{2.33}$. These divisions are shown on the graph sheet for each plotted point (curve 1). They could have been obtained directly by measuring the deviations from curve 2 in percentage with dividers (only on log paper); in practice they would be obtained this way.
3. Use the dividends obtained in step 2 as the third variable.
4. Proceed with the graphical multiple regression as described previously.

The triangular symbols near curve 1 are the points adjusted for the effect of Q_{av}. The fact that they show less scatter than the original points indicates that estimates of Q_{100} are improved by using Q_{av} as an additional variable. It can be shown by computing the equation of the graphical relation that it is of the form

$$\log Q_{100}=\log a+b_1 \log Q_{2.33}-b_2 \log Q_{av}.$$

The introduction of the dividend is merely an expedient; it cancels out of the final relation.

Choice of graphical or analytical method for multiple regression

A standard graphical method is particularly useful for exploratory work and for making preliminary estimates. The graphical method has the following advantages:

1. It is rapid.
2. It helps define the appropriate model.
3. It points out the need for transformations, if any.
4. It brings attention to extremely wild points if they exist in the data. (See the wild point in fig. 18.)

Disadvantages of a graphical method are:

1. Small effects of independent variables cannot be identified.
2. The number of independent variables is limited to about three because of the cumulative effect of inaccuracies in plotting and in locating the lines.
3. Tests of significance of the effects of individual variables are not available.
4. The resulting relation involving three or more variables is confusing to the user unless expressed mathematically or replotted in another form.

An analytical method has the following advantages:

1. For the model used, it gives the best estimate

of the equation constants, and of the standard error.

2. It allows testing of the coefficients for significant difference from zero.
3. Results can be presented in a clear, concise manner which most hydrologists can under stand.
4. Results are unique for the model and sample used; different investigators would get the same results.

Disadvantages of an analytical method are:

1. Computation is time consuming, especially for several variables and complicated models, use of computers reduces the actual computation time but requires considerable time to prepare the data.
2. The existence of wild points is masked as would be the existence of a group of points much different from the majority (unless departures of all points from the estimates are computed).
3. The model selected may not be the appropriate one.

In general a graphical method should be used for exploratory work and the final conclusions should be based on a computed relation.

Determining Equations of Graphical Relations

Graphical analyses are often adequate for certain problems. The results may be reported by furnishing copies of the graphs, but interpretation of graphs in more than two variables is difficult for someone not familiar with the procedure. For instance, consider the three-variable relation of figure 18. What is the expected runoff corresponding to a water content of snow of 20 inches and a precipitation at Three Creek of 10 inches? It is 40,000 acre-feet from the left curve plus 14,000 from the right curve, a total of 54,000 acre-feet. A better method of presentation would be as a family of curves. Another way would be to write the equation of the graphical relation. The equation for the relation of figure 18 is

$$R=-22+1.6S+4.4P \qquad \begin{array}{l} 12<S<28 \\ 4<P<11. \end{array}$$

The limits of definition to the right of the equation tell the reader that he applies the equation to values of S and P outside those limits at his own risk.

Another advantage of defining the equations of graphical relations appears when it is desired to compare relations of the same type but developed from different data. For example, Riggs (1965) related base flow discharges of nine small streams to drainage area and percentage of basin cleared. He made eight different relations, each based on measurements of the same streams but at different times. Interest was in the variability of the effect of the percentage of basin cleared; this variability was apparent when the equations of the relations were defined and the regression coefficients of the percentage of basin cleared were compared.

Still another use for the equation of a graphical relation is to reduce a relation to its simplest terms. This reduction is a desirable procedure if the graphical analysis uses compound variables. The graphical regression of figure 20 is the result of an exploratory study of data for a basin in Western United States and indicates that MAF may be estimated quite reliably from drainage area and mean flow in cubic feet per second per square mile. Because drainage area is used twice, the actual effect of drainage area should be assessed. We begin by writing the equation of the relation (by a method described subsequently), which is

$$\log MAF=1.00+\log A+1.02 \log \overline{Q},$$

where A is drainage area in square miles and $\overline{Q}$ is mean flow in cubic feet per second per square mile. Let $\overline{q}$ be mean flow in cubic feet per second. Then

$$\overline{Q}=\overline{q}/A.$$

Substituting this for $\overline{Q}$ in the first equation gives

$$\begin{aligned}\log MAF&=1.00+\log A+1.02 \log (\overline{q}/A)\\ &=1.00+\log A+1.02 \log \overline{q}-1.02 \log A.\end{aligned}$$

Thus the net regression coefficient of log A is -0.02 which is negligible and, if eliminated from the relation, leaves

$$\log MAF=1.00+1.02 \log \overline{q},$$

or

$$MAF=10\overline{q}^{1.02}.$$

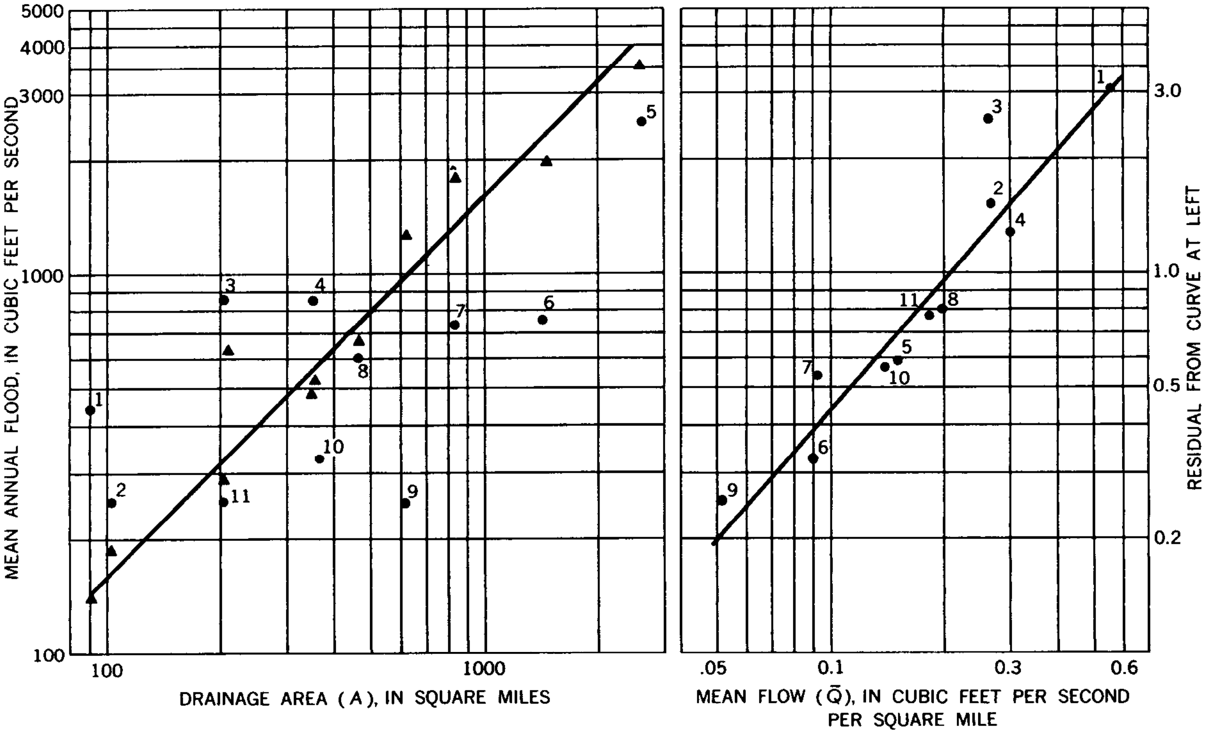

Figure 20.—Graphical regression in which one variable is used twice.

General methods

All linear equations in two variables are of the form

$$Y=a+bX,$$

and this general form is the equation of a straight line on rectangular graph paper. The linear form on log paper is

$$\log Y=\log a+b \log X,$$

which, when expressed in the original variables, is the power equation

$$Y=aX^b.$$

A straight line on semilog paper has the linear form

$$\log Y=\log a+bX,$$

which reduces to the exponential equation,

$$Y=a(10)^{bX}.$$

If $b=c \log k$ in the above equation, then

$$Y=ak^{cX}.$$

Occasionally, points plotted on log paper define a gentle curve rather than a straight line. The locus of the points can sometimes be made linear by adding or subtracting a constant from one of the variables. The relation would be of the form

$$\log Y=\log a+b \log (X+c),$$

or

$$Y=a(X+c)^b.$$

To determine the equation of any linear two-variable relation, compute the slope of the line, b, as vertical distance divided by horizontal distance. These distances are always measured in arithmetic units even though the plot is on log paper (the b values are not transformed). The scale interval should be the same on both axes or an appropriate arithmetic adjustment made. The intercept, a, is usually read off the graph sheet on the ordinate scale at the appropriate value of X. For the relation

$$Y=a+bX,$$

$$Y=a \text{ when } X=0,$$

and for the relation

$$\log Y=\log a+b \log X,$$

$$Y=a \text{ when } X=1.$$

If the graphical line cannot be conveniently extended to $X=1$ or $X=0$, the coordinates of a point on the curve can be substituted in the equation and the intercept can be computed.

The standard equations given in analytic geometry texts are of little use in empirical analysis. More flexible mathematical expressions are needed, and ones that may be put in linear form are desirable because of ease in computing the equation. If a transformation cannot be found that will make the relation linear, then a model of the type

$$Y=a+b_1X+b_2X^2+b_3X^3+\ .\ .\ .\ +b_nX^n,$$

or some portion of it, will fit most plotted smooth curves. If the curvature is only in one direction, the X^2 term will introduce the needed curvature. For a curve having a point of inflection, both the X^2 and the X^3 terms are needed. Terms having higher exponents are rarely used in empirical work.

The above model is equally applicable where X is replaced by log X. A line curved in one direction on log paper is expressed by

$$\log Y=\log z+b_1 \log X+b_2 (\log X)^2.$$

Reducing this to power form gives

$$Y=aX^{b_1}X^{b_2\ \log X}=aX^{b_1+b_2\ \log X}.$$

The general form of linear relations in several variables is

$$Y=a+b_1X_1+b_2X_2\ .\ .\ .\ +b_nX_n.$$

Sometimes the regression coefficient for one variable changes with another variable. This is known as an interaction. In the model

$$Y=a+b_1X_1+b_2X_2+b_3X_1X_2,$$

the last term is called a product interaction term. Its use provides a systematic change in slope.

Curvilinear relations in several variables may be described by adding terms in powers of the independent variables. The equation of a curved line or of a multiple relation involving an interaction is not easily computed. The advantage of recognizing the general form of equation which would represent a particular graphical relation lies in the need for a model if a least-square regression is to be computed. The definition of the equation of a graphical regression is limited to linear regressions.

Definition of equations

The methods for defining the equation of a graphical regression will be demonstrated by two examples. The procedures used in these examples can be adapted readily to other problems. The first example, shown in figure 21, is a multiple linear regression by the method of residuals. The equation of this relation is obtained as follows. Consider first the relation between Y_c and X_1 where Y_c is the curve value from the left part of figure 21. This relation is of the form $Y=a+bX_1$ where a is the intercept at $X_1=0$ and b is the slope of the line. For this example,

$$Y_c=5.4+0.86X_1.$$

The equation of the second line is obtained similarly and is

$$\text{Residual}=-10+2.78X_2.$$

The residual (call it R) is the individual point value, Y, minus the value obtained from the first equation; that is,

$$R=Y-Y_c=-10+2.78X_2.$$

Substituting for Y_c in the above equation gives

$$Y-(5.4+0.86X_1)=-10+2.78X_2$$

from which the desired relation,

$$Y=-4.6+0.86X_1+2.78X_2,$$

is obtained.

The second example (fig. 22) is a relatively simple coaxial graphical multiple regression adapted from one made by the Hydraulic Research Branch of the Bureau of Public Roads. This regression is linear, and the lines for S and P are systematically spaced and parallel. Under these conditions the equation of the graphical relation can be determined.

The following facts are evident from a study of figure 22:

1. Q_{10} is the dependent variable.
2. A is the principal independent variable.
3. Lines of equal P are linearly spaced on logarithmic paper and are parallel.

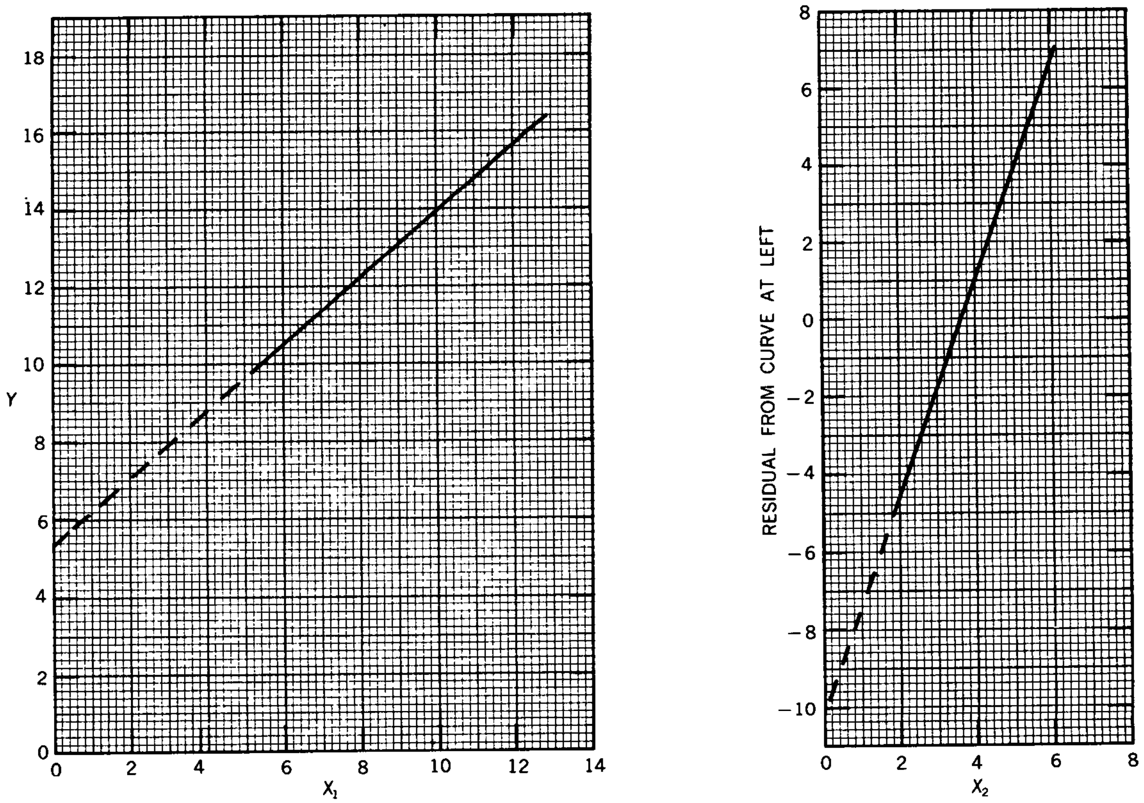

Figure ºº1.—Multiple linear regression by the method of residuals.

4. Lines of equal S are logarithmically spaced at twice the logarithmic scale of the paper and are parallel.

To solve, separate the regression into two parts by introducing an intermediate variable Q_{adj} (to an arbitrary scale) so that

$$Q_{adj}=f(A, P)$$

and

$$Q_{adj}=f(S, Q_{10}).$$

Consider the first relation. For a fixed P, the model would be

$$Q_{adj}=KA^{n},$$

in which K is the intercept on the Q_{adj} scale (at $A=1$) and n is the slope of the line. In this example, $n=1.28$, which is the ratio of the linear vertical to horizontal lengths. When $P=1.20$, the intercept K is 78. To obtain this intercept graphically requires a long curve extension. It is simpler to compute the intercept from some other value of A than one. For instance, for $Q_{adj}=1,000$, $A=7.3$. Then

$$1,000=K(7.3)^{1.28}, \text{ from which } K=78.$$

Similarly, for $Q_{adj}=10,000$, $A=44.5$ and $K=78$.

Introducing P as a variable makes it necessary to define the intercept K in terms of P (because the intercept is different for each value of P). The interval per tenth difference in P when projected on the Q_{adj} scale is 1.36, that is, for each tenth increase in P, the intercept increases 1.36 times (the intercept is on a logarithmic scale). This increase can be measured for individual intervals or computed from the total increase. For instance, for $A=10$:

$$Q_{adj}=1,480 \text{ for } P=1.2,$$

and

$$Q_{adj}=17,000 \text{ for } P=2.0.$$

The increase for eight intervals is 17,000/1,480=11.5, and $(1.356)^8=11.5$.

Then

$$K=78(1.36)^{10(P-1.2)},$$

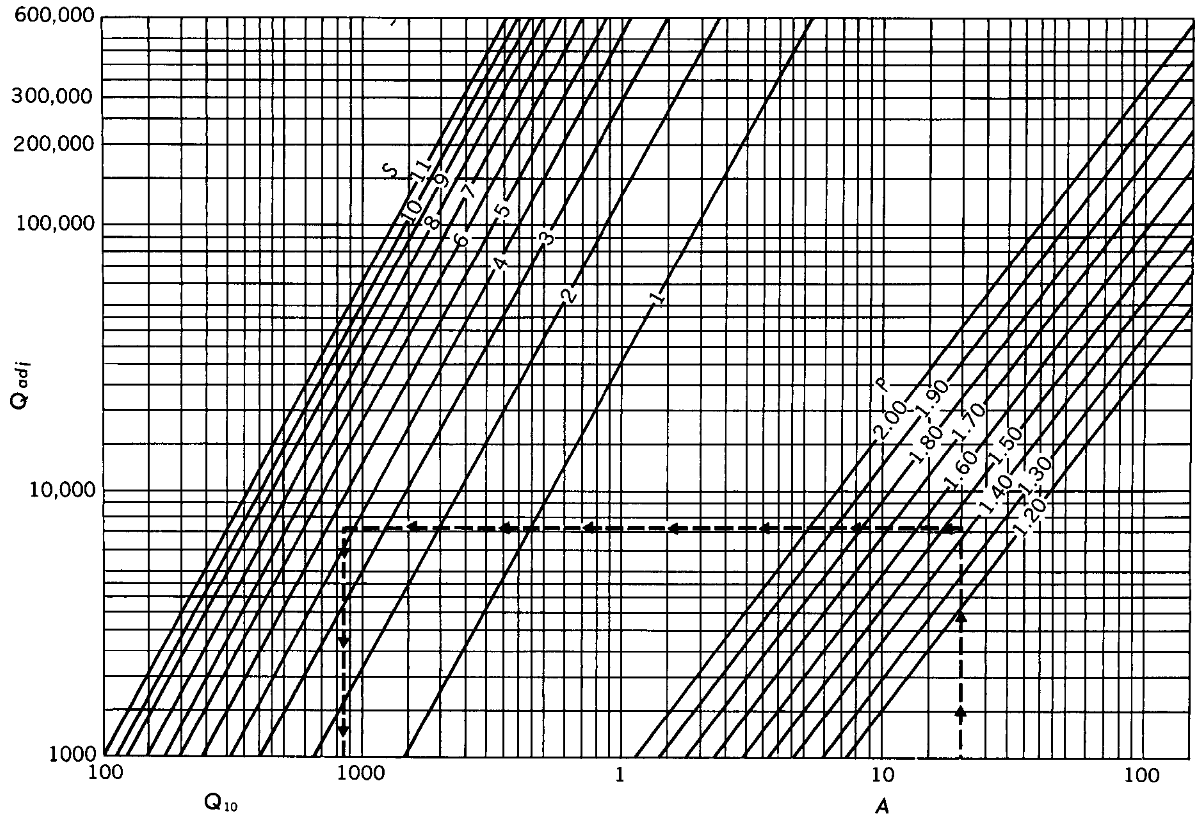

Figure 22.—Coaxial graphical multiple regression.

in which 78 is the intercept at $P=1.2$, the increase in K per tenth is 1.36, and the factor $10(P-1.2)$ is the number of tenths above 1.2.

Substituting the values of K and n gives, for the first relation

$$Q_{adj}=78(1.36)^{10(P-1.2)}A^{1.28}. \qquad (1)$$

The second relation,

$$Q_{adj}=f(S,\ Q_{10})$$

is handled by considering again that Q_{adj} is the dependent variable. Neglecting S, the model would be

$$Q_{adj}=KQ_{10}^{m}.$$

The slope m is 1.78 by scaling. The intercept K is computed at $S=1$, $Q_{adj}=100{,}000$, and $Q_{10}=19{,}000$, as follows:

$$100{,}000=K(19{,}000)^{1.78},$$

or

$$\log 100{,}000=\log K+1.78 \log 19{,}000,$$

$$\log K=5-7.60=-2.60=7.40-10,$$

$$K=0.0025.$$

The spacing of the lines in a vertical direction (parallel to the Q_{adj} scale) is twice the paper scale. Therefore, the intercept K varies directly as S^2, and the equation is

$$Q_{adj}=0.0025S^2Q_{10}^{1.78}. \qquad (2)$$

Equating equations 1 and 2 gives

$$0.0025S^2Q_{10}^{1.78}=78(1.36)^{10(P-1.2)}A^{1.28},$$

which, in logarithmic form, is

$$\log 0.0025+2 \log S+1.78 \log Q_{10}=\log 78$$
$$+10(P-1.2) \log 1.36+1.28 \log A.$$

Solving for Q_{10} gives

$$\log Q_{10}=1.61+0.72 \log A-1.12 \log S+0.75P,$$

which is the desired result.

Other Tools

Analysis of variance

Analysis of variance is a procedure by which the variation embodied in the data of the sample may be resolved into component variations due to independent factors. It is closely related to correlation but is applicable to problems where some of the factors can be described only by classes, not as numerical variates.

The analysis depends on the additive characteristic of variances. Its purpose is to test whether several means are alike or not. The basic features of the process are (1) the measurement of variance among experimental data by the sum of the squared deviations of the observations from their mean, (2) the partitioning of the total sum of squared deviations into independent parts, each part associated with some physical feature of the experiment, (3) the estimation of parameters in the distributions postulated to underlie the data, and (4) tests of significance regarding these parameters. Results of the test give the probability of there being a significant difference between the effects of a factor or factors at different levels.

A very simple example of an analysis of variance concerns whether the mean runoffs for two periods of record at a gaging station are estimates of the same population mean. The annual runoffs are given below:

Period 1	*Period 2*
17.3	6. 4
21.9	15. 2
13.6	9. 7
10.8	4. 4
19.7	9. 9
20.7	11. 9
16.3	11. 9
16.2	15. 4
12.5	9. 4
11.3	7. 0
14.0	16. 0
16.5	17. 0
15.3	11. 2
19.2	13. 2
13.0	11. 5
238.3	170. 1

Computations are as follows:

$$\text{Grand total}=T=238.3+170.1=408.4.$$

$$\text{Total number of items}=N=30.$$

$$\text{Number of items in each period}=n=15.$$

$$T^2/N=(408.4)^2/30=5{,}559.7.$$

Sum of squares of all individuals

$$=\sum Y_{ia}^2=6{,}067.3.$$

$$(\text{Sum of squares of sums})/n=\sum T_i^2/n$$
$$=[(238.3)^2+(170.1)^2]/15=5{,}714.7.$$

$$\text{Between-periods sum of squares}=\sum T_i^2/n$$
$$-T^2/N=5{,}714.7-5{,}559.7=155.0.$$

$$\text{Within-periods sum of squares}=\sum Y_i^2$$
$$-\sum T_i^2/n=6{,}067.3-5{,}714.7=352.6.$$

$$\text{Total sum of squares}=\sum Y_{ia}^2-T^2/N=6{,}067.3$$
$$-5{,}559.7=507.6.$$

The analysis of variance table is

Source	Sum of squares	Degrees of freedom	Mean square	Average mean square
Between periods	155. 0	1	**155	$\sigma^2+n\ \sigma_\xi^2$
Within periods	352. 6	28	12. 6	σ^2
Total	507. 6	29	------	---------

**Statistical significance above the 0.01 level.

The degrees of freedom, D.F., are one less than the number of periods, p, for the between-periods sum of squares and $N-1$ for the total. Thus the degrees of freedom associated with the within-periods source is $N-p$. Mean square is obtained by dividing the sum of squares by D.F.

The last column in the analysis of variance table shows expected values of the mean squares. If the means for the periods are alike, the term $n\sigma_\xi^2$ would be zero. Estimates of the ratio $[\sigma^2+n\sigma_\xi^2]/\sigma^2$ may be greater than one, because of chance or because there is a real difference. This ratio has the F distribution and can be tested statistically. The ratio in the above table is 155/12.6=12.3. The value of F for 1 and 28 degrees of freedom and a probability of 0.01 is $F_{1,28,0.01}=7.6$ from a table of F distribution. Because the sample ratio exceeds the tabular ratio, we conclude that there is a real difference between periods; that is, the probability is less than 0.01 that such a difference in means would have occurred by chance if there were no real difference between periods. The double asterisk on the mean square for between periods (in the analysis-of-variance table) denotes statistical significance above the 0.01 level.

Now consider a similar problem, to determine whether mean annual precipitation at three stations is different. The data are given in the following table.

Year	Precipitation, in inches, at—		
	Site 1	Site 2	Site 3
1945	40.6	48.2	47.5
1946	36.1	40.2	34.8
1947	37.5	37.8	42.2
1948	52.3	58.2	59.9
1949	42.2	43.3	51.7
1950	40.6	41.4	42.5
1951	38.3	42.3	40.5
1952	45.8	48.2	47.8
Sums	333.4	359.6	366.9
$\bar{Y}$	41.7	45.0	45.9

From the data in the table we can make the following calculations:

$$T=1,059.9$$
$$T^2/N=46,807.8$$
$$\Sigma Y^2_{ia}=47,784.0$$
$$\Sigma T_i^2/n=46,885.4$$

The analysis-of-variance table is

Source	Sum of squares	Degrees of freedom	Mean square
Among sites	77.6	2	38.8
Within sites	898.6	21	42.9
Total	976.2	23	---------

$F_{2,21}=38.8/42.9<1$; therefore there is no difference statistically among the three means.

A perusal of hydrologic literature will turn up very few applications of analysis of variance. Most analyses of variance are based on data from a designed experiment, and it is this aplication for which the best results are obtained. Hydrologic data are usually parts of a time series which may not be stationary. Thus the individual values may not be entirely independent as required for a valid analysis of variance. In the example comparing mean runoffs for two periods of record, it was concluded that a real difference existed between periods. But there is no physical reason to expect a change in this basin. The earlier period was one of high precipitation; the later period included the drought of the thirties. It is also possible that some of the annual runoffs were serially correlated. Thus the characteristics of the data tend to discredit the results of this particular application of the analysis of variance.

In the last example, the precipitation at site 2 is greater than that at site 1 for every year shown, yet the analysis of variance shows no difference in means. (An analysis of variance between site 1 and site 2, only, shows a difference at a probability level of about 0.25.) The annual precipitations at a site may be independent, but the precipitations at the several sites for the same year are not. Therefore the requirements of the method are not met, and the results must be accepted with reservation.

The two examples given utilize a very simple statistical model. For more complicated problems, several models may be considered. Selection of the appropriate one is difficult for the "part-time" statistician. Many statistics texts treat analysis of variance in detail. See Bennett and Franklin (1954), Brownlee (1960), and Dixon and Massey (1957). In general, an analysis of variance made by someone not thoroughly familiar with the process should be reviewed by a statistician for suitability of the model and correctness of the interpretation.

Analysis of covariance

The analysis of variance of thė runoff data for two periods (see p. 32) indicated that the population means were probably different, yet other information, particularly precipitation records, leads to the opposite conclusion. The precipitation data can be incorporated in the analysis by using an analysis of covariance. This method includes concepts from analysis of variance and regression and is applicable where a variable represents a measurement for each individual as opposed to a variable which can only be separated into a few categories.

Two general conditions for which an analysis of covariance will produce conclusions different from an analysis of variance are shown in figure 23. In each condition, Y is the variable being analyzed and X is the independent variable. Plot A of figure 23 shows means of Y for the two periods to be practically equal. For this condition an analysis of variance would show no significant difference between means. But a major change in the relation of Y to X occurred between periods 1 and 2, and it is this change that the analysis of covariance can identify.

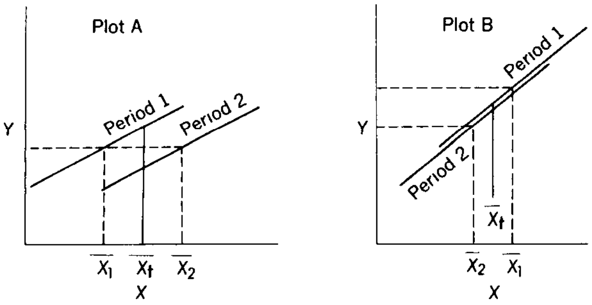

Figure 23.—Two conditions for which analysis of covariance will produce conclusions different from those of analysis of variance.

The analysis of covariance test is made on deviations from regression rather than on means. The test involves the sum of squares of deviations from a regression defined by all points plotted about their own period means and the sum of squares of deviations from an overall regression line (Dixon and Massey, 1957, p. 210). In effect the test indicates whether the two periods are different when adjusted to the same X value. As previously stated, an analysis of variance of data of the condition of plot A, figure 23, would indicate no difference between periods because the means $\bar{Y}_1$ and $\bar{Y}_2$ are nearly alike. But analysis of covariance would show a significant difference in Y values corresponding to the overall mean $\bar{X}_t$.

Plot B of figure 23 shows two periods having very different mean Y values but no real difference in the regressions of Y on X for the two periods. An analysis of variance would show a significant difference between means, but an analysis of covariance would show no significant difference in regressions for the two periods. The two results do not conflict. There is a difference in means for the two periods, but this difference is due to a difference in X values for those periods.

Analysis of covariance requires that slopes of the regression lines for the individual periods be virtually parallel. A test for parallelism has been described by Dixon and Massey (1957, p. 218).

Table 5.—Annual precipitation index and annual runoff, for example of analysis of covariance

Period 1		Period 2	
Precipitation index (X)	Runoff (Y)	Precipitation index (X)	Runoff (Y)
27	17. 3	14	6. 4
36	21. 9	26	15. 2
26	13. 6	15	9. 7
18	10. 8	11	4. 4
27	19. 7	19	9. 9
30	20. 7	21	11. 9
25	16. 3	18	11. 9
28	16. 2	22	15. 4
19	12. 5	20	9. 4
22	11. 3	17	7. 0
22	14. 0	29	16. 0
29	16. 5	30	17. 0
26	15. 3	16	11. 2
29	19. 2	23	13. 2
24	13. 0	23	11. 5
388	238. 3	304	170. 1

$T_x=692$; $T_y=408.4$.

Details of an analysis-of-covariance computation are given below using (1) the same runoff data as in the previous section for the analysis-of-variance example and (2) some assumed values of a precipitation index, all of which are listed in table 5 and plotted on figure 24. The plot indicates that there is no change in

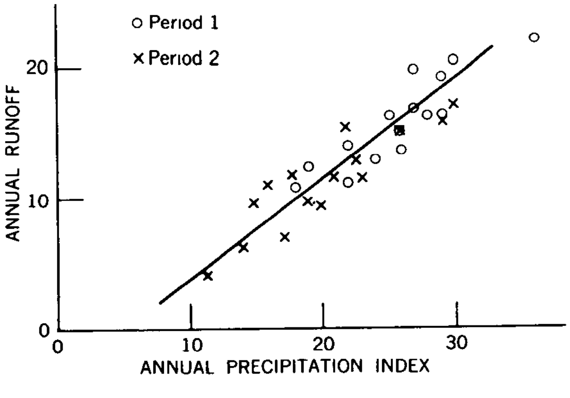

Figure 24.—Plot of data from table 5.

relation between periods and that the difference in runoff means between periods was due to differences in precipitation. For these data it would not be necessary to make a covariance analysis. However, if separate regressions were indicated by the plotted points, it might be desirable to make a covariance analysis in order to test whether the two regressions were significantly different statistically.

The following computation illustrates the procedure:

Total sum of products$=\sum X_{ij}Y_{ij}-T_xT_y/nk$, where T_x and T_y are grand totals of X and Y, n is the number of items in each period, and k is the number of periods.

Between-means sum of products

$$=\sum T_{xi}T_{yi}/n-T_xT_y/nk,$$

where T_{xi} and T_{yi} are column (period) totals. For this example,

the total sum of products$=(27)(17.3)$

$$+(36)(21.9)....+(23)(11.5)$$

$$-(692)(408.4)/(15)(2)$$

$$=10{,}054.7-9{,}420.5=634.3$$

Between-means sum of products

$$=(388)(328.3)/15+(304)(170.1)/15$$

$$-(692)(408.4)/30=9{,}611.4$$

$$-9{,}420.4=191.0$$

Total sum of squares on $X=\sum X^2{}_{ta}$

$$-T_x^2/N=16{,}898-15{,}962=936$$

Between-periods sum of squares on X

$$=\sum T_{xi}^2/n-T_x^2/N=16{,}197-15{,}962=235$$

Sums of squares on Y are taken from the analysis-of-variance example.

Deviations from regression are computed by the formula

$$\sum y^2-(\sum xy)^2/\sum x^2,$$

which, for totals, $=507.6-(634.3)^2/936=507.6-429.8=77.8$.

For within periods, the deviations from regression are

$352.6-(443.3)^2/701=352.6-280.3=72.3$,

and the between-means deviation from regression is obtained by subtraction. The data are shown in the following covariance table.

Source	Data				Deviations		
	Degrees of freedom	Σx^2	Σxy	Σy^2	Degrees of freedom	$\Sigma(Y-\overline{Y})^2$	Mean square
Between means	1	235	191	155.0	1	5.5	5.5
Within periods	28	701	443.3	352.6	27	72.3	2.7
Total	29	936	634.3	507.6	28	77.8	-------

Sums of squares for within periods (in the first part of the table) are obtained by subtraction. Degrees of freedom for deviations from regression, $\Sigma(Y-\overline{Y})^2$, for within periods and total are one less than for means.

The test of significance compares F (the ratio of mean squares from the covariance table), to values of the distribution of F at the 5-percent and 10-percent levels. For this example they are

$$F=5.5/2.7=2.0,$$

$$F_{1,27,0.05}=4.2,$$

and

$$F_{1,27,0.10}=2.9.$$

Because 2.0 is less than 2.9, the difference in periods is not significant at the 10-percent level when runoffs are adjusted for precipitation.

See the article by Wilm (1943), which includes a discussion by Davenport, for an application of covariance analysis to a hydrologic problem.

Multivariate analysis

Multiple regression on independent variables which are related among themselves sometimes produces inconsistent results from different sets of data. For example, the regression coefficient of an independent variable may range from positive to negative in different regressions and yet test statistically significant in each. Under these conditions, the conclusions regarding the effect of that variable on the dependent variable might be wrong if only one set of data was analyzed. The use of multivariate analysis has been proposed as a way out of this dilemma.

Multivariate analysis is concerned with the relationship of sets of dependent variates and includes several different procedures, each intended to accomplish a different objective. Snyder (1962) investigated the use of multivariate analysis in hydrology where the structure of the solution was of primary interest. Kendall (1957) described the theory.

In its present (1965) state of development, multivariate analysis is not a useful tool for defining cause-and-effect relationships in hydrology; regression analysis is still the best method available.

Characteristics of Hydrologic Data

Streamflow is a continuous process which varies with time, and thus streamflow data are said to form a time series. A plot of streamflow against time would show a pattern of variation recurring each year; that is, high flows tend to occur at particular times of the year and low flows at others in response to climatological characteristics which also vary seasonally.

Because streamflows are not discrete values, we need to chop the hydrograph into pieces which we will consider as individual streamflows. The particular pieces we use have certain characteristics which must be considered in analysis. The most common piece is the daily mean discharge. A daily mean discharge is related to the discharge of the previous day and lies within a range which depends on the time of year. In statistical terms, daily mean discharge is a serially correlated variable, that is, it is nonrandom. The daily mean discharges for a year are also not homogeneous; they are more likely to be larger at one time of the year than at another. Data are considered homogeneous if any subgroup to which certain of these data may be logically assigned has the same expected mean and variance as any other subgroup of the population.

Monthly mean discharges for different calendar months are also serially correlated and nonhomogeneous. Annual mean discharges may be homogeneous values. They may or may not be serially correlated, depending on the amount of basin storage at the time that the hydrologic year begins.

Instead of a streamflow variable made up of adjacent segments of a hydrograph, we may consider variables such as July mean, annual peak discharge, or annual minimum flow. These variables are made up of one individual from each year and thus are independent of the yearly cycle of streamflow. They are also independent of each other (with the possible exception of annual minimum flows which include effluent from ground-water recharge of a previous year).

Precipitation, temperature, sediment discharge, water quality, transpiration, evaporation, and solar radiation vary throughout the year; indices describing them may be nonrandom and nonhomogeneous.

Obviously the distinction between random and nonrandom data and between homogeneous and nonhomogeneous data is not always clear cut. The analyst will have to determine whether the effects of possible moderate nonrandomness or nonhomogeneity will invalidate the conclusions of his particular analysis. It is important that the character of the data be considered in designing the analysis and in interpreting the results.

So far we have described variables that may be considered samples from a population if the individuals are homogeneous. If the individuals are also random, we can estimate the frequency distribution of the variable from the sample. Another type of variable used extensively in hydrology cannot be considered to have a probability distribution, or even to be drawn from a population as thought of in the usual sense. Basin characteristics such as drainage area, slope, elevation, and vegetal index are in this category. (It is possible to conceive of certain physiographic parameters as random variables, but rarely can the available sample be considered randomly selected or representative.)

Time is sometimes used as a variable in regression. It has no distribution and is used only as a substitute for the real factor or factors (which are unknown or cannot be expressed by indices) associated with changes in a dependent variable.

Effects of data characteristics on analysis

We prepare a frequency distribution of daily mean discharges from several years of data; this is the duration curve. The individual values are nonrandom and nonhomogeneous. Therefore the duration curve cannot be considered a frequency curve. The probability of exceeding a certain value on a particular future day depends both on the preceding value and on the time of year. Thus, the duration curve is merely the distribution of daily means that has occurred. It can be considered an estimate of the distribution during a future period several years long.

On the other hand, frequency curves of annual flood peaks can be interpreted as probability curves because the individuals are unrelated and homogeneous. Most low-flow frequency curves can be similarly interpreted, but occasionally a serially correlated sample will be found.

The effect of using nonhomogeneous data in a regression problem is shown by figure 25 in which is plotted 4 years of monthly mean discharge for each of the 12 calendar months for two stations, one in Turkey and one in Idaho. The relation looks fairly good, but there is actually no relation between the two streams for a particular calendar month. The apparent relation using all calendar months arises because the yearly cycle of streamflow in Idaho resembles that in Turkey. Discharges in winter months are low and in spring snowmelt months are high.

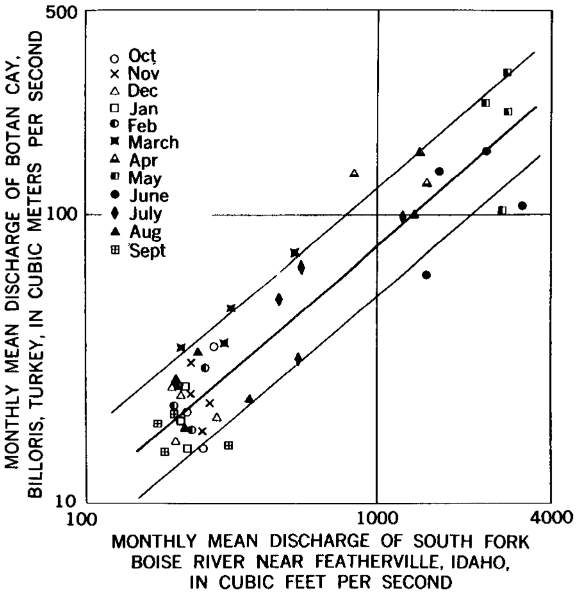

Figure 25.—Spurious relation using nonhomogeneous data.

Less extreme conditions are shown by relations between monthly mean discharges from contiguous basins. For example, there is no relation between monthly mean discharges of Lake Fork above Moon Lake, Utah, and Duchesne River at Provo River Trail, Utah, for the calendar month of January; there is a fair relation for the calendar month of June (fig. 26). With few exceptions the relation between monthly discharges from two adjacent drainage basins for a particular calendar month is not the same as the relation for a different calendar month. When monthly discharges for all calendar months are used together, the computed correlation coefficient will be too high and the computed standard error will be an average of the standard errors for the individual calendar-month relations.

Outliers

Many factors influence the flow of a stream; some exert great influence at one time and none at another; most exert effects which are interrelated with effects of other factors. Only a few factors can be included in a regression used to estimate streamflow, and the effects of these factors are only approximated. Consequently there is a scatter of points about the regression line and occasionally a wild point occurs (see fig. 18 for an example). Such wild points are called outliers in statistics, and statistical tests are available for use in determining whether or not a particular point should be rejected as not belonging to the group. It seems questionable whether outliers in hydrologic analyses should be rejected on the basis of a statistical test. Consider the wild point in figure 18. If the precipitation had been about 7 inches instead of 3.4 inches, the point would not be wild. It is possible that precipitation at the higher elevations in the Jarbidge River basin was much greater than at Three Creek. If it were, the same thing could happen again and more weight should have been given to that point in the analysis. However, if some of the data for that year are found to be unreliable we could reject the point.

Acton (1959) devoted a short chapter to the rejection of unwanted data. He says, in part, "But the plain truth is that physical scientists and engineers need not be encouraged to ignore obstinate outlying data—rather they need to be held in check".

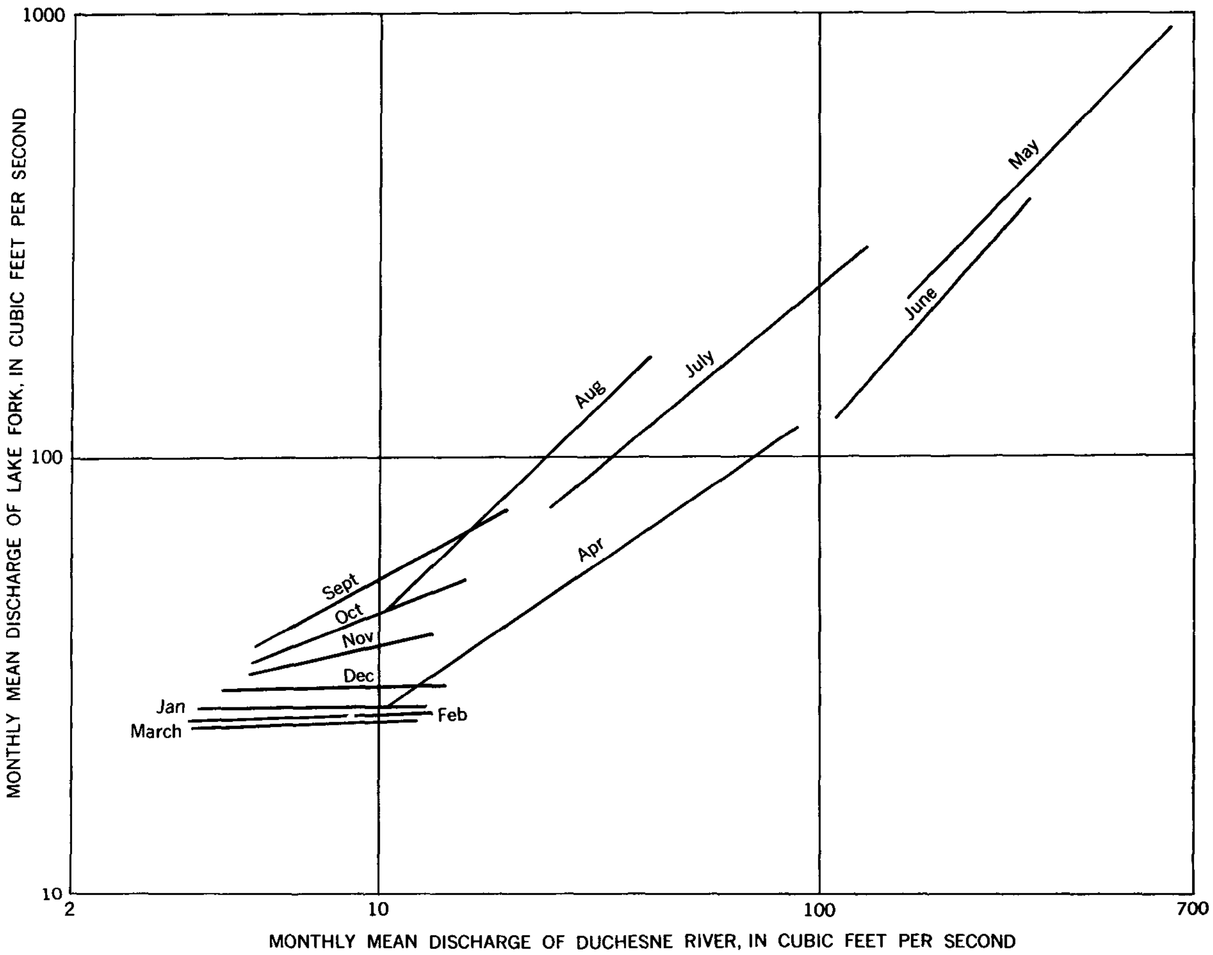

Figure 26.—Discharge relations for individual months, two Utah stations.

Selected References

Acton, F. S., 1959, Analysis of straight-line data: New York, John Wiley & Sons, Inc., 265 p.

Amorocho, J. and Hart, W. E., 1964, A critique of current methods in hydrologic systems investigation: Trans. Am. Geophys. Union, v. 45, no. 2, p. 307–321.

Bennett, C. A., and Franklin, N. L., 1954, Statistical methods in chemistry and the chemical industry: New York, John Wiley & Sons, Inc., 724 p.

Benson, M. A., 1960, Characteristics of frequency curves based on a theoretical 1,000-year record, *in* Dalrymple, Tate, Flood-frequency analysis: U.S. Geol. Survey Water-Supply Paper 1543–A, p. 51–74.

——— 1962, Factors influencing the occurrence of floods in a humid region of diverse terrain: U.S. Geol. Survey Water-Supply Paper 1580–B, 64 p.

——— 1965, Spurious correlation in hydraulics and hydrology: Am. Soc. Civil Engineers Proc., v. 91, no. HY4, p. 35–42.

Brownlee, K. A., 1960, Statistical theory and methodology in science and engineering: New York John Wiley & Sons, Inc., 570 p.

Dawdy, D. R., and Matalas, N. C., 1964, Analysis of variance, covariance, and time series, *in* Chow, V. T., Handbook of applied hydrology: New York, McGraw-Hill Book Co., p. 8–68 to 8–90.

Dixon, W. J., and Massey, F. J., 1957, Introduction to statistical analysis: New York, McGraw-Hill Book Co., Inc., 488 p.

Ezekiel, M., 1950, Methods of correlation analysis: 2d ed., New York, John Wiley & Sons, Inc., 531 p.

Ezekiel, M., and Fox, K. A., 1959, Methods of correlation and regression analysis: New York, John Wiley & Sons, Inc., 548 p.

Fisher, R. A., 1950, Statistical methods for research workers: New York, Hafner Publishing Co., 355 p.

Gumbel, E. J., 1958, Statistics of extremes: New York, Columbia Univ. Press, 371 p.

Hazen, Allen, 1930, Flood flows: New York, John Wiley & Sons, Inc., 199 p.

Kendall, M. G., 1952, The advanced theory of statistics: London, Charles Griffin and Co., v. 1, 457 p.

——— 1957, A course in multivariate analysis: London, Charles Griffin and Co., Ltd., 185 p.

Linsley, R. K., Kohler, M. A., and Paulhus, J. L. H., 1949, Applied hydrology: New York, McGraw-Hill Book Co., Inc., 689 p.

McDonald, J. E., 1957, A critical evaluation of correlation methods in climatology and hydrology: Arizona Univ. Inst. Atmospheric Physics Sci. Rept. 4, 35 p.

Mood, A. M., 1950, Introduction to the theory of statistics: New York, McGraw-Hill Book Co., Inc., 433 p.

Riggs, H. C., 1958, Discussion of paper by E. Kuiper, "100 frequency curves of North American rivers": Am. Soc. Civil Engineers Proc., v. 84, no. HY1, paper 1558, p. 61–63.

——— 1960, Discussion of paper by A. L. Sharp, A. E. Gibbs, W. J. Owen, and B. Harris, "Application of the multiple regression approach in evaluating parameters affecting water yields of river basins": Jour. Geophys. Research, v. 65, no. 10, p. 3509–3511.

——— 1965, Effect of land use on low flows in Rappahannock County, Virginia *in* Geological Survey Research 1965, U.S. Geol. Survey Prof. Paper 525–C, p. C196–C198.

Siegel, Sidney, 1956, Nonparametric statistics: New York, McGraw-Hill Book Co., Inc., 312 p.

Snedecor, G. W., 1948, Statistical methods: 4th ed., Iowa State Coll. Press, 485 p.

Snyder, W. M., 1962, Some possibilities for multivariate analysis in hydrologic studies: Jour. Geophys. Research, v. 67, no. 2, p. 721–729.

U.S. Geological Survey, 1949, Floods of August 1940 in the Southeastern States: U.S. Geol. Survey Water-Supply Paper 1066, 554 p.

Wilm, H. G., 1943, Statistical control of hydrologic data from experimental watersheds: Am. Geophys. Union Trans., 1943, pt. 2, p. 618–624 [with discussion by Davenport].

☆ U.S. Government Printing Office: 1977—240-961/123